TAMPERE UNIVERSITY OF TECHNOLOGY

HUY TRAN

SEQUENCE-DEPENDENT NOISE FILTERING IN SMALL GENETIC MOTIFS

Master of Science Thesis

Examiners: Andre S. Ribeiro
Olli Yli-Harja
Examiners and subject approved in the Faculty of Computing and Electrical Engineering Council meeting on 6.3.2013

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Information Technology
HUY TRAN: Sequence-dependent Noise Filtering in Small Genetic Motifs
Master of Science Thesis, 44 pages
May 2013
Major: Signal Processing
Examiners: Andre S. Ribeiro, Olli Yli-Harja
Supervisor: Antti Häkkinen
Keywords: Gene Expression, Motif, Filter, Toggle switch, Stochastic, Noise, Simulation, Promoter, Sequence, Synthetic Biology

The rates of intracellular processes are, in general, in constant change in response to environmental signals and other internal processes. To deal with noise in the input signals, filtering and decision-making circuits are needed. Motivated by recent evidences from *in vivo* measurements that the rate limiting steps in transcription initiation are critical in determine RNA and protein numbers, we study the effects of these steps on the behavior of three genetic circuits: a toggle switch, a genetic amplitude filter and a genetic frequency filter. We model these circuits, and from stochastic simulations, we study the performance of the filters and the stability of the switch. We find that these features degrade as the transcript levels are lowered. These effects can be alleviated by adding rate limiting steps to the transcription initiation process. In addition, we show that some features of the filters, such as cutoff levels, are affected by changes in mRNA production dynamics as well. In conclusion, our study shows that the kinetics of transcription initiation of the genes composing these circuits, which are largely determined by the promoter sequence, can be varied within realistic parameter ranges of values to alter considerably their behaviors.

PREFACE

This content of the thesis is based on my research with Antti Häkkinen and Andre Ribeiro on the effect of promoter kinetics on the behavior of small genetic circuits, at the Laboratory of BioSystem Dynamics (LBD), Department of Signal Processing, Tampere University of Technology.

I would like to thank Antti Häkkinen for his guidance and supervision over my work since my early acquaintance with the field of computational biology, and Jason Lloyd-Price for his user-friendly stochastic simulator, SGNSim.

I would also like to extend my gratitude to Asst. Prof. Andre S. Ribeiro who has not lost his patience in guiding me as an IT student closer to the perspective of biologists and physicists. He also improved the writing of the thesis by giving valuable comments, without which this work would not have been possible.

Tampere, May 15th, 2013

Huy Tran

CONTENTS

ABBREVIATIONS AND SYMBOLS

E.coli	Escherichia coli
CME	Chemical Master Equation
DNA	Deoxyribonucleic Acid
DSSA	Delayed Stochastic Simulation Algorithm
mRNA	Messenger Ribonucleic Acid
NRM	Next Reaction Method
ODE	Ordinary Differential Equation
RNA	Ribonucleic Acid
RNAP	Ribonucleic Acid Polymerase
RBS	Ribosome Binding Site
SGNS	Stochastic Genetic Networks Simulator
SSA	Stochastic Simulation Algorithm
TSS	Transcription Starting Site

1. INTRODUCTION

All living organisms ensure their existence by maintaining gene expression, the synthesis of mRNA and consequently proteins from the information stored in their DNA. The synthesis of the first product is called transcription. In transcription, the RNA polymerase binds to the promoter region of the gene and trigger the transcription initiation process to set the promoter to the initiation state. The RNA polymerase then begins to elongate along the gene sequence and assemble mRNA. When a termination site on the DNA is recognized, the elongation is terminated with the release of RNA polymerase and mRNA from the promoter.

Recent evidences suggest that, in organisms such as bacteria, the kinetics of transcription initiation is the key in determining the expression patterns of genes. Once the promoter closed complex is formed, the RNA polymerase remains at the promoter sequence, preventing the next transcription initiation event to occur until the open complex is complete[26]. This is a multi-step process[6] that takes a long time and, thus, plays an important role in determining the mean and noise in mRNA and consequently protein number. The number and duration of the steps vary between promoters, even between those with small differences in their sequences[24]. Recent *in vivo* measurements[18] showed a sub-Poisson production of mRNA when the gene is induced at weak or medium levels, indicating that the process of transcription initiation of the promoter involves a sequential mechanism with at least two elementary steps. The duration of these steps can vary with induction, temperature, etc.

Cells' functions are carried by groups of genes. In each of these, the genes interact with each other through their proteins to create genetic motifs, which are capable of more complex patterns of behavior than individual genes. Different topologies result in different dynamics[41]. Several motifs have been identified. Examples include genetic switches, which can function as memory circuits; genetic clocks, which can be used for time keeping and synchronization purposes; and genetic filters, which can be used for noise filtering or computation.

It is unknown to what extent the promoter dynamics, i.e. the kinetics of transcription initiation, affects the behavior of the cellular motifs. Here, we examine how these kinetics affect the functionality of motifs performing noise filtering and decision making, which are critical for cells to survive in fluctuating environments. As many essential genes' products, especially those with signaling functions, exist

in low copy numbers, the stochasticity in the molecules' kinetics cannot be ignored in the dynamics of motifs. It is necessary to assess its effect on the mean behavior of the motifs. To capture both the low copy number effect and the stochastic nature of gene expression, we utilize a stochastic approach to model processes. We model sequence-dependent noise filtering and decision making genetic circuits by modeling explicitly the time that each process in transcription initiation takes to be completed once initiated, rather than modeling the process itself. Thus, the sequence-dependence of the process is represented in the distribution of the intervals between transcription events.

In the next chapter, the biological background as well as the modeling of the genetic circuits are provided. We employ two approaches in modeling the intracellular processes: the deterministic approach to predict the dynamics of the chemical systems qualitatively, and the stochastic approach to assess the system in consideration of molecules' stochastic kinetics and low copy number effects. We present the Stochastic Simulation Algorithm (SSA)[12] through which, we are able to obtain the systems' realizations with the probability in accordance with the Chemical Master Equation[14]. Next, we show the model of single gene expression which captures the sequential-step transcription initiation's kinetics, and the regulation of one gene to another via its products. Based on these models, we build the models of genetic circuits as chemical systems. On the account of the promoter delay during transcription initiation, we use the Delayed Stochastic Simulation Algorithm (DSSA)[3], derived from SSA, to realize the dynamics of the motifs.

The third chapter describes our results on the performance of filtering and decision-making systems with different kinetics in transcription initiation. First, we investigate the effects of rate-limiting steps in transcription initiation on the expression dynamics of individual genes. We study the behavior of three circuits: a toggle switch[1] with the decision-making function, and two genetic filters in the amplitude and frequency domain. For the decision-making system, we evaluate how long the switch takes to change from one state to another, and the distribution of the intervals between switching events. For genetic filters, we are interested in the locations of pass-bands and stop-bands, the quality (e.g, signal attenuation at the pass-bands, the steepness of the transition bands) and the noise level of the output signal. The response of the circuits is quantified for a wide range of transcriptional dynamics that are in accordance with measurements[39, 18]. The simulations are generated by the Stochastic Genetic Network Simulator (SGNSim) by Andre S. Ribeiro and Jason Lloyd-Price[33]. The simulation of the deterministic models and the analysis of the simulation data are performed using MATLAB [25].

The fourth and final chapter concludes the thesis with a discussion on the relevance of the results, and with suggestions on the construction of synthetic circuits

with arbitrary connections.

2. BACKGROUND

2.1 Chemical kinetics in cells

In this section, we first described the means by which chemical reactions are represented, and how these are utilized to describe elementary biological processes in cells. Next, we present two models of these processes, one deterministic and the other stochastic. They provide us with platforms with which we can simulate and inspect the dynamics of chemical systems.

2.1.1 Representation of chemical reactions

Chemical reactions can be qualitatively represented by reaction equations, which consist of reactants, products and arrows (usually read as 'yields') indicating the transformation direction of reactants toward products. Below are typical examples of reaction equations:

$$\emptyset \rightarrow A \tag{2.1}$$

$$A \rightarrow \emptyset \tag{2.2}$$

$$A + B \leftrightarrow AB \tag{2.3}$$

$$AB \rightarrow AC \tag{2.4}$$

$$AC \rightarrow A + C \tag{2.5}$$

$$B + A \rightarrow B + C \tag{2.6}$$

The reaction equations 2.1 to 2.6 involve 3 substances: A, B and C. Reaction 2.1 describes the creation of one molecule of substance A, whereas reaction 2.2 indicates its degradation. Reaction 2.3 describes the formation of complex molecule AB from one molecule of A and one molecule of B. The bidirectional arrow indicates that this reaction is reversible: the complex molecule AB can dissipate into separate molecules of A and B. Reaction 2.4 describes the transformation of complex molecule AB into AC and reaction 2.5 describes the breakdown of AC into molecules of A and C. If we put equations from 2.3 to 2.5 into one system, the overall trend is the transformation of substance A to C with B acting as catalyst, which is equivalent to equation 2.6.

This qualitative representation of all reactions in a system gives us an overview of the factors affecting reactants' dynamics, but fails to predict what the global

trends are. In the case of more complex systems that involve many reactions and substances, we usually cannot track the change in substances' molecule numbers or concentration over time.

2.1.2 Deterministic models

Each reaction has its own kinetics, which represented typically by a reaction rate equation. It links the reaction rate with reactants' concentration and stoichiometric coefficients to determine how frequent a specific chemical reaction occurs. Consider a generic chemical equation with no intermediate steps (also called elementary reaction):

$$aA + bB \rightarrow cC \tag{2.7}$$

For this reaction, we define the rate equation as follow:

$$r = k[A]^a[B]^b \tag{2.8}$$

whereas $[A]$ and $[B]$ are the concentration of species, and usually measured in moles per liter (called molarity or M). a and b are the stoichiometric coefficients of the balanced equation. They indicate how many molecules of each reactant are consumed for the reaction to occur. k is the rate coefficient or rate constant of the reaction. The value of k depends on the characteristic of the reactant molecules and the environment conditions such as temperature, surface area, ionic strength between molecules, ect. r is the reaction rate of the elementary reaction. The value of r is proportional to the product of the reactants' concentration and the rate constant, following the law of mass action. We add k above the arrow of chemical reaction:

$$aA + bB \xrightarrow{k} cC \tag{2.9}$$

The total number of molecules required for one reaction to occur $a + b$ is the order of the chemical reaction. A reaction can be zeroth-order, where the concentrations of the reactants are constant (e.g, a substance is produced from very stable source, at a constant rate). In practice, an elementary reaction should not have an order higher than two, as the chance for three molecules to collide and trigger reaction is minute. Therefore, it should be decomposed into lower order elementary reactions.

Next, we inspect the dynamics of a solution with N species from X_1 to X_N and

M reactions: the j^{th} reaction is described as follow:

$$\sum_{i=1}^{N} s_{ij} X_i \xrightarrow{k_j} \sum_{i=1}^{N} r_{ij} X_i \tag{2.10}$$

k_j is the rate constant of the j^{th} reaction. s_{ij} and r_{ij} indicates how many molecules of substance X_i are consumed/produced via j^{th} reaction. Effectively, there are $S_{ij} = r_{ij} - s_{ij}$ molecules of X_i created after each j^{th} reaction occurs. S_{ij} are elements of the matrix S, size of N by M, called the stoichiometric matrix.

The reaction rates are calculated as follow:

$$r_j = \sum_{i=1}^{N} [X_i]^{s_{ij}} \tag{2.11}$$

r_j constitute the reaction rate vector r, length of M. Note that stoichiometric coefficients s_{ij}, r_{ij} are not necessarily positive integers. Leaving them zero indicates the absence of the species in the reaction. S_{ij} can take positive (corresponding to production of X_i via j^{th} reaction), negative (degradation of X_i via j^{th} reaction) or zero (no changes) values.

$[X_i(t)]$ denotes the concentration of species X_i at time t. We are interested in the change in the concentration vector function $[X(t)]$ to fully describe the kinetics of the chemical system. Assuming that the solution's volume goes to infinity, the number of reactant molecules also goes to infinity, if maintaining finite species concentrations. $r(t)$ is the value of reaction rate vector at time t, calculated from the vectors $[X(t)]$ and k. For each reaction, we calculate the yielding rate of each species X_i as follow:

$$[\text{yielding rate of } X_i] = [\text{reaction rate}] \times [\text{number of } X_i \text{ molecules produced}]$$

The total yielding rate of a species in the system equals the sum of yielding rates in all reactions. Thus, we derive a formula that deterministically characterizes the changes in [X(t)] over time $\partial[X(t)]/\partial t$:

$$\frac{\partial[X(t)]}{\partial t} = S.r(t) = f([X(t)]) \tag{2.12}$$

The general equation 2.12 is also called the Ordinary Differential Equation (ODE). The derivatives are ordinary because they only apply to functions of independent variables. The partial derivative value of a species' concentration at a specific time can take any real value, either positive, negative or zero. The set of differential equations $\partial[X(t)]/\partial t = f([X(t)])$ allow us to study the time evolution of $[X(t)]$ as its solution.

We take an example of a simple system with a zeroth-order protein production

(2.13) and a first order protein degradation (2.14).

$$\emptyset \xrightarrow{k_P} \tag{2.13}$$

$$P \xrightarrow{d_P} \emptyset \tag{2.14}$$

The protein molecule is denoted by P, and the protein concentration in cell is [P]. From equation 2.12, we have:

$$\frac{\partial[P(t)]}{\partial t} = k_P - d_P[P(t)] \tag{2.15}$$

Given at time t_0, there is no protein in the system, the protein concentration over time has the solution:

$$[P(t)] = \frac{k_P}{d_P}(1 - e^{-(t-t_0)\times d_P}) \tag{2.16}$$

It is noted that the system reaches the stationary phase when $[P(t)] = k_M/d_M$, corresponding to $\partial[P(t)]/\partial t = 0$. If the initial protein concentration is different from this limit, it will take an infinitely long time for the concentration to reach this stationary phase.

The deterministic model may give us some information on the dynamics trend of molecule numbers in the system, and particularly useful when assessing processes with a population approach. Nevertheless, cells' volume is limited within the membrane. Some essential genes' products, especially ones with signaling functions, exist in cell only in low copy number. Their concentrations, the unit of which is molecules over liter, are therefore discrete and not uniform from cells to cells. The deterministic time evolution of substances' quantities may differ greatly from practice. Additionally, we cannot evaluate the fluctuations in the substance quantities arise from the random movement of molecules in solution, which is ignored when employing the deterministic ODE model.

2.1.3 Stochastic models

In this section, we present a model that captures both the effects from low copy numbers and the stochastic nature of chemical reactions. First, we study the kinetics of molecules in a solution based on the following assumptions:

- The system is well-stirred and of constant volume V.
- The system is in thermal equilibrium at constant temperature T.
- Reactions occur only when two or more molecules collide, while most collisions do not lead to reactions.

The first assumption requires that the spatial distribution of one species' molecules is uniform within the volume V. Also, the position of molecules is independent of each other, no matter whether they are of the same species or not. The second assumption indicates every molecule in the solution has independent, normally-distributed velocity, with mean equal k_BT/m. k_B is the Boltzmann constant[20], T is the solution's absolute temperature and m is mass of molecule. The third assumption ensures the system follows Maxwell-Boltzmann statistics, when the temperature is high enough but reactant density (excluding the fluid density) is low enough[20]. These assumptions allow us to ignore individual molecules' position and velocity, and adopt a probabilistic approach. Instead of predicting the exact molecule number of each substance $x_i(t)$, we find their probability density $\pi(x_i)(t)$.

The stochastic form for equation 2.10 is therefore written as follow:

$$\sum_{i=1}^{N} s_{ij} X_i \xrightarrow{c_j} \sum_{i=1}^{N} r_{ij} X_i \tag{2.17}$$

The reaction constant c_j is "reaction probability per time unit" instead of "reaction rate", indicating how likely the reaction j^{th} is to happen given the reactants' molecule number at a given time. Its value can be calculated from the collision rate, Maxwell's velocity, and the reaction probability when collisions between reactants happen. The propensity function, instead of rate equation, is defined as:

$$a_j(\mathbf{x}) = c_j.h(\mathbf{x}) \tag{2.18}$$

$a_j(\mathbf{x})dt$ indicates the probability for the j^{th} reaction to occur in the infinitesimal time window $[t, t+dt)$. $h(\mathbf{x})$ is the number of possible reactant combinations of a reaction at a given time, with $\mathbf{x}$ the reactants' molecule number vector. Because the propensity function at a specific time depends only on the current state rather than previous ones, we can consider the system dynamics as a Markov process, where each reaction marks a change in state.

Assuming the state probability of the system at time t: $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, we can calculate $P(\mathbf{x}, t+dt|\mathbf{x}_0, t_0)$ the probability of the system in state $\mathbf{x}$ at time $t+dt$ from propensity functions:

$$\begin{aligned} P(\mathbf{x}, t+dt|\mathbf{x}_0, t_0) = P(\mathbf{x}, t|\mathbf{x}_0, t_0).P(no\ reaction\ in\ [t, t+dt)) + \\ \sum_{j=1}^{M} (\mathbf{x} - S_j, t|\mathbf{x}_0, t_0).P(one\ reaction\ j^{th}\ in\ [t, t+dt)) \end{aligned} \tag{2.19}$$

When there is no reaction in the window $[t, t+dt)$, the system's state does not change. When one among M reactions occurs, the new probability is updated with

the probability of the previous state $P(\mathbf{x} - S_j, t|\mathbf{x}_0, t_0)$ and its propensity function. We find that:

$$P(one\ reaction\ j^{th}\ in[t, t+dt)) = a_j(\mathbf{x} - S_j)dt \tag{2.20}$$

$$P(no\ reaction\ in[t, t+dt)) = (1 - \sum_{j=1}^{M} a_j(\mathbf{x}))dt \tag{2.21}$$

From equations 2.20 and 2.21, we can rewrite equation 2.19 as follow:

$$P(\mathbf{x}, t+dt|\mathbf{x}_0, t_0) = P(\mathbf{x}, t|\mathbf{x}_0, t_0).(1 - \sum_{j=1}^{M} a_j(\mathbf{x}))dt + \sum_{j=1}^{M} (\mathbf{x} - S_j, t|\mathbf{x}_0, t_0).a_j(\mathbf{x} - S_j)dt \tag{2.22}$$

The system dynamics therefore is described by the taking the first derivative of equation 2.22:

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \lim_{dt \to 0} \frac{P(\mathbf{x}, t+dt|\mathbf{x}_0, t_0) - P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{dt} \tag{2.23}$$
$$= \lim_{dt \to 0} \frac{\sum_{j=1}^{M} (\mathbf{x} - S_j, t|\mathbf{x}_0, t_0).a_j(\mathbf{x} - S_j)dt - P(\mathbf{x}, t|\mathbf{x}_0, t_0).(\sum_{j=1}^{M} a_j(\mathbf{x}))dt}{dt}$$

Equation 2.23 is commonly known as the first order Chemical Master Equation (CME)[14], describing the time-evolution of probability of a chemical system to occupy each one of the discrete space of states. It provides the basis to find the state probability of a system at a specific time given its initial conditions, which is usually intractable using analytical or numerical approaches[13].

It was observed and proved mathematically that with very large number of reactants' molecules and reaction volume, the stochastic model's results converge to that of the deterministic model[21]. However, when the molecule numbers are limited, the stochastic results, even in mean, stray far from the deterministic ones. In these results, not only the trends but also fluctuations in reactant number are considered, thus making the stochastic approach preferable when evaluating a chemical system.

Cell is, however, far from a typical chemical system. First, it is based overwhelmingly on carbon compounds: from such macro-molecules as DNA, mRNA, proteins, lipids, to simple ones as carbohydrates, vitamins[5]. Second, it is highly complex, involving zillions of molecules, capable of interacting with each other to form the metabolic network. Any biological processes though simplest cannot be treated as one but a sequence of elementary reactions. With the stochastic approach, though offering more information on the system dynamics (e.g fluctuations), an appropriate

modeling strategy is required to gain results of relevance and avoid the explosion of the parameter range.

2.2 Stochastic simulation

2.2.1 Stochastic simulation algorithm

Obtaining an analytical solution to Chemical Master Equation (CME) is computational challenging and usually intractable especially for complex systems involving a large number of substances. The generation of the probability densities is calculated on the continuous time scale, which is vulnerable to error when performed on a digital computer. The Stochastic Simulation Algorithm (SSA)[12, 14] is employed to address the problems.

The SSA numerically simulates the underlying Markov process that the CME describes using random sampling. A single simulation of the system over time executes explicitly a single possible sequence of reactions, yielding a single realization in the possible state space of the system with the appropriate probability density. The acquisition of the result is simple and generally inexpensive with current digital computers. However, to characterize the system dynamics, one is usually interested in the probability density described by the CME, thus, multiple realizations are generated and analyzed together to approximate the real density. The creation of multiple trajectories to yield an exact estimation can become expensive in time and data storage. Nevertheless, SSA is still a preferred approach which allows the user to allocate resource consumption to achieve a certain level of approximation.

Another reason that supports the preference of SSA is that, by studying the dynamics of substances in single trajectories, one can infer information about the trends as well as dependence of substances' quantity unobtainable by CME. For example, robust oscillation with fluctuation in its period, when averaged, become damped oscillation. The information about the mean and noise in period is therefore not extractable.

Unlike CME which tracts the time evolution of the probability density at fixed rates, SSA realizes the time evolution of substance number with random rates drawn from a distribution. The distribution, which describes the density that the next reaction in the system is the j^{th} reaction and will occur in an infinitesimal time interval $[t+\tau, t+\tau+d\tau)$, is dependent of the temporal system's state $\mathbf{x}$ at time t. We divide the interval $[t, t+\tau)$ into k smaller windows of size $\epsilon = \tau/k$. It is required that no reactions occur in any of these windows, and the j^{th} reaction occurs in the

time window $[t+\tau, t+\tau+d\tau)$. From equation 2.20 and 2.21, we have the distribution:

$$P_{\tau,j}(\tau, j; \mathbf{x}, t)d\tau = (1 - \sum_{j=1}^{M} a_j(\mathbf{x})\epsilon)^k (a_j(\mathbf{x})d\tau) \tag{2.24}$$

We denote $a(x) = \sum_{j=1}^{M} a_j(x)$ the sum of all reactions' fluxes. Dividing the equation sides by $d\tau$, taking the limit $d\tau \to 0$, we have:

$$\begin{aligned} P_{\tau,j}(\tau, j; \mathbf{x}, t) &= (1 - a(\mathbf{x})\epsilon)^k a_j(\mathbf{x}) \\ &= (1 - a(\mathbf{x})\tau/k)^k a_j(\mathbf{x}) \end{aligned} \tag{2.25}$$

which takes the limit when $k \to \infty$:

$$\begin{aligned} P_{\tau,j}(\tau, j; \mathbf{x}, t) &= a_j(\mathbf{x})\text{exp}(-a(\mathbf{x})\tau) \\ &= a'_j(\mathbf{x})a(\mathbf{x})\text{exp}(-a(\mathbf{x})\tau) \end{aligned} \tag{2.26}$$

where $a'_j(\mathbf{x}) = a_j(\mathbf{x})/a(\mathbf{x})$ is the normalized flux of reaction j^{th}. The term $a(\mathbf{x})\text{exp}(-a(\mathbf{x})\tau)$ represents the exponential distribution of the probability of one reaction occurs at time $t + \tau$. The normalized flux $a'_j(\mathbf{x})$ indicates how likely it is the j^{th} reaction.

Equation 2.26 is the basis for the realization of SSA: for any state $\mathbf{x}$ in the system's state space, we can calculate the time τ for the next reaction to occurs by inversing the distribution $a(\mathbf{x})\text{exp}(-a(\mathbf{x})\tau)$, and identify this reaction from $a'_j(\mathbf{x})$. The inverse transformation is performed using a pair of uniform random numbers $r_1, r_2 \sim U[0, 1)$. We can draw τ and j from the following equations:

$$\tau = -a(\mathbf{x})^{-1} ln(1 - r_1) \tag{2.27}$$

$$j = j' \text{such that} \sum_{i=1}^{j'-1} a'_i(\mathbf{x}) \leq r_2 < \sum_{i=1}^{j'} a'_i(\mathbf{x}) \tag{2.28}$$

We have the outline to implement the algorithm. From any given state $(\mathbf{x}, t)$, we calculate τ, j from (x) and update the system with the new state $\mathbf{x} + S_j$ at time $t+\tau$. S_j is the stoichiometric vector indicating the changes in molecule numbers after one reaction j^{th} occurs. Therefore, from the initial state $(\mathbf{x}_0, t_0)$, we can generate one trajectory describing the time evolution of substances' molecule numbers. The algorithm is executed in the following steps:

1. Initialize the step $n = 0$, time: $t_n \leftarrow t_0$ and state: $\mathbf{x}_n \leftarrow x_0$
2. Calculate $a_j(\mathbf{x}_n)$ and $a(\mathbf{x}_n)$ from the current state $\mathbf{x}_n$
3. Generate r_1,r_2 from uniform distribution [0,1)

4. Calculate τ and j from equations 2.27 and 2.28

5. Perform reaction R_j: Update $t_{n+1} = t_n + \tau$, $\mathbf{x}_{n+1} = \mathbf{x}_n + S_j$

6. Set $n = n + 1$. Return to steps 2

During the simulation based on the algorithms, the record of all $(\mathbf{x}_n, t_n)$ is maintained, giving a complete realization of the system' state dynamics described by CME. The trajectory is usually re-sampled at fixed time intervals to make the analysis of multiple trajectories more convenient.

2.2.2 Delayed stochastic simulation algorithm

One drawback of SSA derived by Gillespie is that it is not possible to model non-elementary (or non-exponential) processes. In the biological context, many complex reactions such as transcription initiation, translation...are sequential, usually time consuming, and thus cannot be ignored. One usually have to break them into or approximate them with sequences of elementary reactions:

$$A \xrightarrow{c_0} I_1 \xrightarrow{c_1} ... \xrightarrow{c_n-1} I_n \xrightarrow{c_n} B \tag{2.29}$$

Such approximation, however, requires to identify the rate of each steps, making the number of free parameters in the model increases drastically, especially for processes with little noise. Furthermore, the increase in number of reactions and reactants also makes the simulation computationally and storage expensive.

A method to address the problem was proposed by Bratsun et. al.[3] and generalized for multiple delayed products by Roussel and Zhu[35]. As the intermediary products I_1 to I_n do not play any other role in the system, reaction 2.29 is shorten to the delayed form:

$$A \xrightarrow{c_0} B(\tau_B) \tag{2.30}$$

The delay term τ_B represents the time elapsed of the transformation from I_1 to B, that is, the molecule B is released at τ_B after the reaction consuming A. τ_B can have arbitrary probability densities, predefined by user rather generated by convolving the densities of elementary reactions. The system is therefore semi-Markovian: the next state does not only depend on the current but also on the previous ones. However, the first reaction rule as applied to the original SSA remains the same, with the release of the delayed product considered one reaction.

The outline of the Delayed SSA (DSSA), derived by extending the SSA is as follow:

1. Initialize the step $n = 0$, time: $t_n \leftarrow t_0$ and state: $\mathbf{x}_n \leftarrow x_0$

2. Calculate $a_j(\mathbf{x}_n)$ and $a(\mathbf{x}_n)$ from the current state $\mathbf{x}_n$

3. Calculate τ and j from equations 2.27 and 2.28

4. If there are delayed products released in the time interval $[t, t+\tau]$:
   (a) Release the delayed product x_i with closest release time t'. Update $t_{n+1} \leftarrow t'$, $x_{n+1(i)} \leftarrow x_{n(i)} + 1$
   (b) Set $n = n + 1$. Return to step 2

5. Perform reaction R_j by updating $t_{n+1} = t_n + \tau$, $textbf{x}_{n+1} = textbf{x}_n + S_j$

6. Set $n = n + 1$. Return to steps 2

In the scope of this thesis, the stochastic simulations are performed with SGNS Stochastic Simulator[33] written by Andre S. Ribero and Jason Lloyd-Price.

2.3 Single gene expression

Upon division, the daughter cells inherit not only the material from their parent but also its instructions to survive. These instructions are stored in a macromolecule called DNA (Deoxyribonucleic Acid) using the language of nucleotides and replicated accurately through cell division.

The central dogma of molecular biology, as claimed by Marshall Nirenberg, can be recapitulated as "DNA makes RNA makes protein". The first process is referred as transcription, and the second translation. The genetic information of DNA is stored in sequences called genes. In transcription, the RNA polymerase (RNAP) transcribes a gene on DNA and assembles a messenger RNA (mRNA) as a single-stranded copy of that gene. In translation, the mature mRNA is bound and read by ribosomes to synthesize proteins or other genetic products. Translation and transcription play a major role in determining the dynamics of gene expression, which, in this context, is described by the protein dynamics. They affect not only the effective rate of protein production but also its fluctuations over time, the cell to cell diversity and thus, cannot be excluded from the model of gene expression.

Since the effect of fluctuations and correlation between mRNA and protein numbers on the protein dynamics is non-negligible, the stochastic chemical kinetics approach is employed. Furthermore, any cellular processes, due to the number and size of involving entities, are complex multi-step. The reading and assembling of mRNA from DNA, for examples, are conducted sequentially on a gene of hundreds to thousands of nucleotides. A model that includes all series of elementary reactions will be over-detailed and become a burden for its computation as well as interpretation. Here, we select from simplified models featuring most time-consuming sub-processes

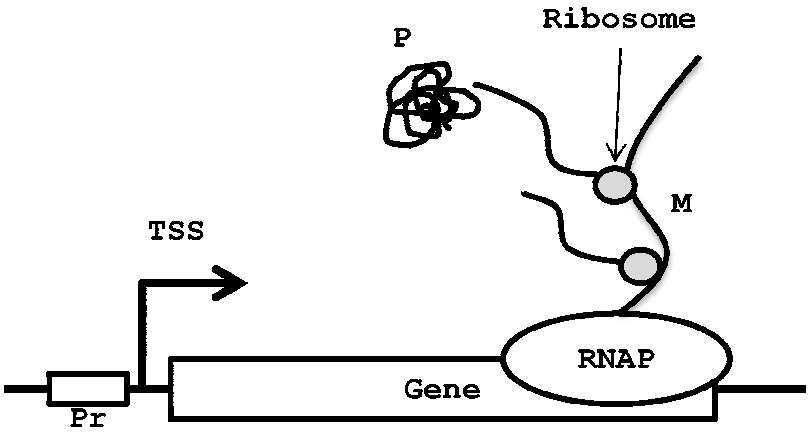

Figure 2.1: **The synthesis of mRNA (M) and protein (P) from DNA:** Transcription is performed by the enzyme RNA polymerase (RNAP) from the Transcription Starting Site (TSS). While mRNA is being elongated, ribosomes bind to the ribosome binding sites (RBS) and translate proteins from the template. The proteins only become functional upon folding.

that capture the variability and correlations in the gene expression on the overall. The series of elementary reactions are replaced with time delayed reactions[11].

Several models to capture the dynamics of gene expression have been proposed, from the simplest one that describes the protein production as a first order reaction[17], single-step transcription-translation[34], to very detailed ones with reactions at nucleotide and cordon levels[31]... To inspect the key stochasticity of gene expression computation, we choose the model of gene expression that separates transcription (2.31) and translation (2.32)[43]:

$$Pr + RNAP \xrightarrow{k_P} Pr(\tau_1) + RNAP(\tau_2) + M(\tau_3) \tag{2.31}$$

$$M + r \xrightarrow{k_P} M + r(\tau_r) + P(\tau_P) \tag{2.32}$$

In reaction equation 2.31, one RNA polymerase, denoted by $RNAP$, binds to the promoter region of the gene, denoted by Pr and initiates transcription. The probability rate constant for the binding is represented by k_M, the product of reaction is one mRNA molecule and the break-down of RNA polymerase - promoter complex. τ_1 denotes the reaction delay of promoter region after transcription initiation. This delay is accounted by several transformations of DNA sequence around the transcription start site, making it available for elongation[9] and then, for new RNA polymerase binding. τ_2 represents the RNA polymerase residence time on the DNA strand. τ_3 is the time elapsed from initiation to mRNA (M) availability for translation. In eukaryote cells, the chromosome is kept in nucleus, where transcription occurs. The completely transcribed mRNA is processed and travels to the rough en-

doplasmic reticulum[9] where ribosomes attach to its ribosome binding sites (RBS) and initiate translation. This extra time is also included in τ_3. Note that the gene in the model exists at single-copy level: there is only one mRNA produced after each transcription event.

Reaction equation 2.32 describes the translation of mRNA by ribosomes (r) to produce proteins (P). τ_r and τ_P denote the delay of ribosome and protein release after the translation initiation. There is no occupied time of mRNA because it has multiple ribosome binding sites, allowing more than one ribosome to bind to simultaneously.

In Prokaryote cells, the lack of nucleus couples transcription and translation[28]. During transcriptional elongation, the unfinished mRNA is already available for ribosome binding and translation. Thus, we consider the mRNA delay no longer than promoter delay:$\tau_1 = \tau_3$. Furthermore, due to RNAPs and ribosomes abundance in cell, we consider their molecule number fixed and omit them in the initiation rates. Finally, we have a simplified model of transcription and translation as follow:

$$Pr \xrightarrow{\infty} Pr(\tau) + M(\tau) \tag{2.33}$$

$$M \xrightarrow{k_P} M + P \tag{2.34}$$

The lifespan of mRNA is generally limited within minutes and fits well with exponential distribution[2]. Proteins with a much longer half life, which can elapse several cell cycles, have their concentration diminished through cell elongation and division. Their decaying process is shown as a first-order reaction[16]:

$$M \xrightarrow{d_M} \emptyset \tag{2.35}$$

$$P \xrightarrow{d_P} \emptyset \tag{2.36}$$

The term decay does not necessarily indicate the disappearance of the molecule in the cell. When a molecule ceases to function properly, is metabolized or transported out of the system, the number of the functional molecules of that species is diminished by one.

By looking at the mechanism how genes express, we find that noise is intrinsic. There are several sources of noise that affect the dynamics of molecule numbers: the kinetics and functional durations of entities, the transcription and translation elapsed time are not deterministic. The resulted fluctuations cannot be hidden due to low copy number effect. A high noise level, though in some extent granting cell flexibility when coping with changes in the environment, reduces cell fitness[40]. A higher variance in gene expression widens the range of protein dynamics: over expression leads to waste of cell materials and energy for unnecessary synthesis,

while the lack of proteins can put several subsequent cellular processes on hold.

2.4 Rate-limiting steps

Because mRNAs exist in cell only at low copy number, gene expression is strongly correlated with the temporal mRNA number, the dynamics of which is determined by the transcription rate and mRNA half-lives[2]. While the degradation rate of mRNAs may regulate the transient changes in abundance in response to environmental stresses, the transcription initiation's dynamics governs the mRNA level in both mean and fluctuation.

In bacteria, transcription initiation at the promoter region is found to comprise sequential rate limiting steps[27]:

$$Pr + RNAP \overset{slow}{\longleftrightarrow} Pr.RNAP \overset{rapid}{\longleftrightarrow} Pr_c \overset{rapid}{\longleftrightarrow} I_1 \tag{2.37}$$
$$\overset{rapid}{\longleftrightarrow} I_2 \overset{rapid}{\longleftrightarrow} I_3 \overset{rapid}{\longleftrightarrow} Pr_o \overset{rapid}{\longrightarrow} Pr_{init}$$
$$Pr_{init} \overset{rapid}{\longrightarrow} Pr + RNAP + M \tag{2.38}$$

The set of reactions 2.37 is the detailed model of the process. First, RNA polymerase diffuses along the DNA template until reaching TSS to form the promoter-polymerase closed complex (Pr_c). This complex undergoes imsomerizations (I_1, I_2 and I_3), which only become rate limiting at low temperature. RNA polymerase then unwinds the DNA strand at TSS to form the open complex (Pr_o)[6, 36]. Once, the promoter is set into initiation state (Pr_{init}), RNA polymerase begins to elongate along the gene sequence and assemble mRNA 2.38. When a termination site on the DNA is recognized, the elongation is terminated with the release of RNA polymerase and mRNA from the promoter.

The durations of these rate-limiting steps vary under different conditions of temperature[27], Mg2+ concentration[19] and other metabolites. *In vivo* measurements showed that the elongation time only last tens of seconds, even in case of transcriptional pauses[15], compared to transcription intervals on the order of thousands seconds. Thus, the delay at TSS is mostly accounted for the intervals between mRNA productions.

To quantify the kinetics of each promoter, we use $\tau \sim \Gamma(\alpha, \lambda^{-1}/\alpha)$, which denotes that the promoter delay τ is drawn from a gamma distribution with the shape of α and rate of $\lambda\alpha$. The total mean duration of all rate-limiting steps is therefore λ^{-1}. In this sense, the transcription initiation comprises α rate-limiting steps, each with the duration of $(\lambda\alpha)^{-1}$. Even if the sequential steps in practice are unequal in duration, the gamma distribution is still a good approximation. If the steps are of the same order of magnitude, they can be considered approximately equal and, else,

fast steps can be neglected. Only the longest steps of the same order are considered significant.

The kinetics of transcription initiation is found to vary from promoters to promoters, even for ones with slight differences in their sequences. In an experiment[24] on a panel of promoter sequences derived from the lac promoter of *E.coli*, the transcription initiation process on these promoters is interrupted at different steps and to various extent. The regulators are found to affect only the first stable complex between RNAP and promoter, while the promoter governs mainly the conversion of the closed complex to the open complex.

2.5 Gene networks

2.5.1 Gene regulation

Gene expression is not spontaneous. Despite RNAP and ribosomes abundance, not all genes on the chromosome are transcribed constantly over time. There are several mechanisms that can affect a gene's output yield or even stop its expression completely. The promoter secondary and ternary structure can be altered, thus restricting its interaction with RNAP and impeding mRNA production. mRNA upon transcribed can be falsely processed or blocked by molecules other than ribosomes, which forbids translation[4]. Those mechanisms grant cells capability to adjust their behavior in respond to internal or external signals. In this thesis, we only discuss the governing of gene expression by transcription factors.

To allow transcription initiation by RNAP, the promoter must be set to a specific state. In this ON state, we say the gene is activated. In the OFF state, the gene is repressed and do not produce mRNA. The state of promoter is decided by operator sites that are located around the promoter region on DNA. Those sites, when interacting with certain molecules, change the structure of promoter along with its state. During transcription initiation, due to the large size of RNA polymerase, it overlaps the the whole promoter protecting the operator state from interactions with other molecules.

While the promoter is not occupied by RNA polymerase, there are several factors capable of specifically binding to the operator region, thus changing the promoter state:

$$Pr + R \xrightarrow{k_{assoc}} Pr.R \tag{2.39}$$

In reaction 2.39, the promoter's operator site interacts with one regulator molecule R forming an operator-regulator complex. This complex may change the promoter structure, block RNAP polymerases physically or actively attract them to the transcription starting site, resulting in the change in promoter rate of transcription ini-

tiation and consequently the gene expression level. If transcription is inhibited upon binding, the regulator is called repressor. If the regulator promulgates transcription, it is called activator.

Reaction 2.39 is reversible: The bound regulator R can escape from the operator region intactly with certain probability:

$$Pr.R \xrightarrow{K.k_{assoc}} Pr + R \tag{2.40}$$

with k_{assoc} the association rate constant in 2.39. K is the disassociation constant indicating the disassociation over association rate ratio. As many regulators are proteins, the lifespan of which is usually much longer (in hours) than the disassociation time ($(K.k)^{-1}$), we ignore the degradation of the binding regulator. Because it is generally not known how the transcription factors regulate the gene expression, we assume that each step is affected in an equivalent manner. The mean duration of the promoter delay is thus dependent of the transcription factors' quantity: $\lambda = k_M f(R)$. k_M is the maximum rate of mRNA production when the gene is fully induced. $f(R)$ is the regulation function determined by the regulator molecule number, taking values from 0 to 1.

To simplify the model, we assume the binding of regulatory proteins is much faster than the rate of transcription, that is $k_{assoc} \to \infty$ while K is kept unchanged. It is clear that the expectation of operator regulator complex $E[Pr.R] = (1+(K.R^{-1})^{-1})$ with given regulator number p. Thus, the regulation function is set as:

$$f(R) = (1 + (K.R^{-1})^{+d})^{-1} \text{ iff R is an activator} \tag{2.41}$$

$$f(R) = (1 + (K.R^{-1})^{-d})^{-1} \text{ iff R is a repressor} \tag{2.42}$$

d denotes the Hill coefficient, determining how steep the transition between the ON and OFF states is. As many proteins are known to function in a dimeric form[42] or regulate genes through cooperative bindings, we use d=2 in all following cases. It is notable that the disassociation constant K is a good indication of the regulatory efficiency of the transcription factor: $f(R = K) = 1/2$. If K is small, the regulation is strong and vice versa.

If the gene is regulated by two regulators A and B, the combined regulation function is the product of two individual regulation functions:

$$f_{AB}(A, B) = f_A(A) f_B(B) \tag{2.43}$$

2.5.2 Motifs

Individual genes have a limited set of possible dynamical behaviors. In practice, genes rarely function separately. By connecting multiple genes' expression together, cell can perform more complex and highly non-linear functions[41]. More complex patterns of behaviors, such as making decisions or counting time, require genetic motifs, which are groups of genes interacting with each other via their products.

If the protein of gene A is gene's B activator, gene A is said to activate gene B. The activation is indicated by a pointing arrow:

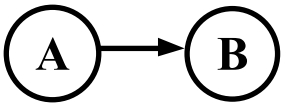

Figure 2.2: **Activation of gene B by gene A.**

Gene A is said to repress gene B if the first's product is the latter's repressor. The repression is represented by a dashed arrow:

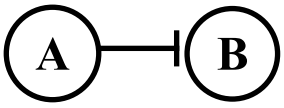

Figure 2.3: **Repression of gene B by gene A.**

A gene can also be self-regulated, that is its product is also its regulator:

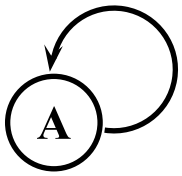

Figure 2.4: **Self activation of gene A.**

The activation or repression of one gene to another is considered a directed connection. The chromosome of simple organisms such as *E.coli* contains thousands of genes with intracellular interactions forming a highly complex regulatory network. The signal can travel from one node to another via multiple overlapping pathways, under different modifications: it can be delayed, relayed (through activation) or reversed (through repression)... As a result, the output gene expression dynamics is far different from that of the input signal. The formation of the gene network with connections gives rise to intracellular complex behaviors, somehow limited by the fluid environment that situates them.

Despite the number of genes and interactions, each function in cell is usually governed by one small group of genes. Those genes interact with each other, forming a small genetic circuit. The circuit's patterns, or motifs, which are repeated throughout the network[29], and determine the expression dynamics of constituent genes[41]. Qualitatively, we find genetic circuits' behavior similar to that of their electrical analog counterparts. Thus, motifs earn the the corresponding names, such as amplitude or frequency filters, clocks, switches...

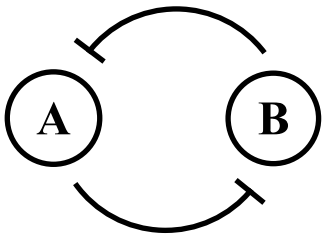

Figure 2.5: **Toggle switch.**

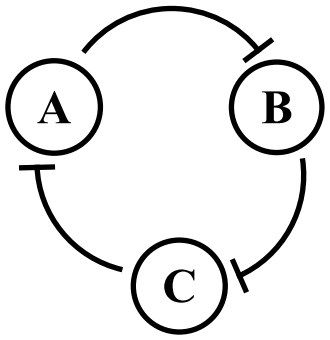

Figure 2.6: **3-gene repressilator.**

In figure 2.5, gene A and gene B strongly repress each other. If gene A is ON, its high protein level keeps gene B OFF, which allows gene A to express even more. On the other hand, if gene A is OFF, gene B is turned ON and becomes the dominating gene. The system is either in one of the two states, and can flip from one state to another randomly. The circuit is thus called toggle switch[1]. Figure 2.6 shows the topology of the repressilator: the system contains 3 genes repressing one another, forming a closed loop. Each gene in the system switches between ON and OFF states successively, leading to the oscillation in its protein level. Though it is unknown whether cell employs the exact patterns in nature, such multi-stability (e.g, the existence of multiple stable fixed points) and clock behavior in gene expression have been observed in a wide range of organisms, from bacteria, plants, and animals.

There have been attempts[10, 8]to construct genetic circuits *in vivo* in E.coli with a certain percentage of the population expressing the desired behavior. The

general strategy to assemble the circuits is to define the connections from the known promoter-regulator pairs. The performance of the circuits is determined not only by the kinetics and efficiency of the regulators but also by the sequence-dependent kinetics of the promoters. In the next chapter, we demonstrate the impact of the promoter kinetics on the behavior and performance of the filtering and decision making circuits.

3. RESULTS

3.1 Single gene expression

We study the effect of rate limiting steps in transcription initiation, determined by the promoter sequence, on the expression of single genes. Namely, we vary the number of steps α and the duration of each steps λ^{-1}; while maintaining the kinetics of other processes; we observe how the level of gene products (mRNA, protein) in mean number and fluctuation changes with different transcription initiation kinetics.

We observe the system behavior in the case of full induction, that is the gene is always on. We assume that the steps in transcription initiation are equal, thus the total promoter delay follows a Gamma distribution $\Gamma(\alpha, \lambda\alpha)$ with the shape of α and mean duration of λ^{-1} (as described in section 2.4). The expected protein level corresponding to the length of those steps is $\mu' = \lambda/d_M k_P/d_P = \lambda/k_M\mu$.

Unless stated otherwise, the parameters of gene expression are set as follow: $d_M = (5min)^{-1}$, $k_M/d_M = 5$, $d_P = 60min^{-1}$, $k_P/d_P = 100$. The expected protein number is $\mu = k_M/d_M k_P/d_P = 500$. The parameters are selected in accordance with *in vivo* measurements in *E-coli.*[39, 2]. For each pair of values (α,λ), we simulate the model in 5×10^6 s using the DSSA. At the start of the simulation, there is no mRNA and protein in the system.

Shown in figure 3.1 is a system trajectory in the first 2×10^4 s, with $alpha = 4$, $\lambda/k_M = 0.05$. Each mRNA degrades before the release of the next one, indicating that the proteins are produced in a bursty manner. Due to the slow protein degradation rate (small d_P), the protein number are still maintained positive over time and loosely correlated with the mRNA production intervals. The protein number is very different from the result of the ODE model, which has the solution $P(t) = \lambda/k_M \times 500 \times exp(-t \times d_P)$.

We inspect the dynamics of molecule numbers when the system is in the steady state. Figure 3.2 shows the protein level with different rate limiting steps. As we increase the duration of mRNA production intervals, the mRNA level decreases along with the protein level. From the stochastic simulation, the measured mean protein matches that of the deterministic model $\mu' = \mu\lambda/k_M = 500 \times \lambda/k_M$.

From the stochastic simulation result, we observe a significant amount of noise in the gene expression, far different from deterministic one (in which the noise level equals zero). Figure 3.3 shows the quantification of noise in mRNA and protein

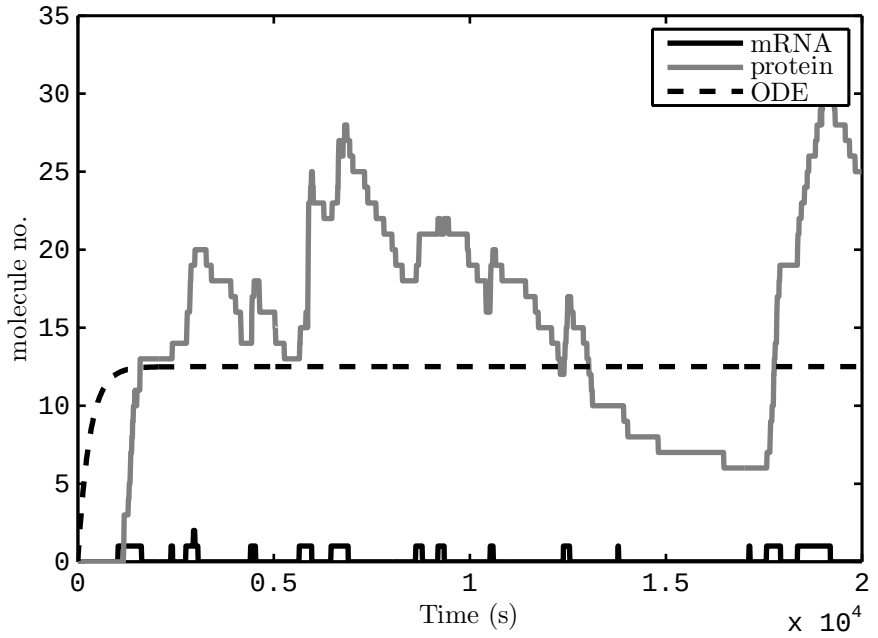

Figure 3.1: **A simulation trajectory of molecules numbers with $\alpha = 4$, $\lambda/k_M = 0.05$:** The time series (solid lines) are shown in comparison with the expected protein number, calculated from the deterministic model (dash line).

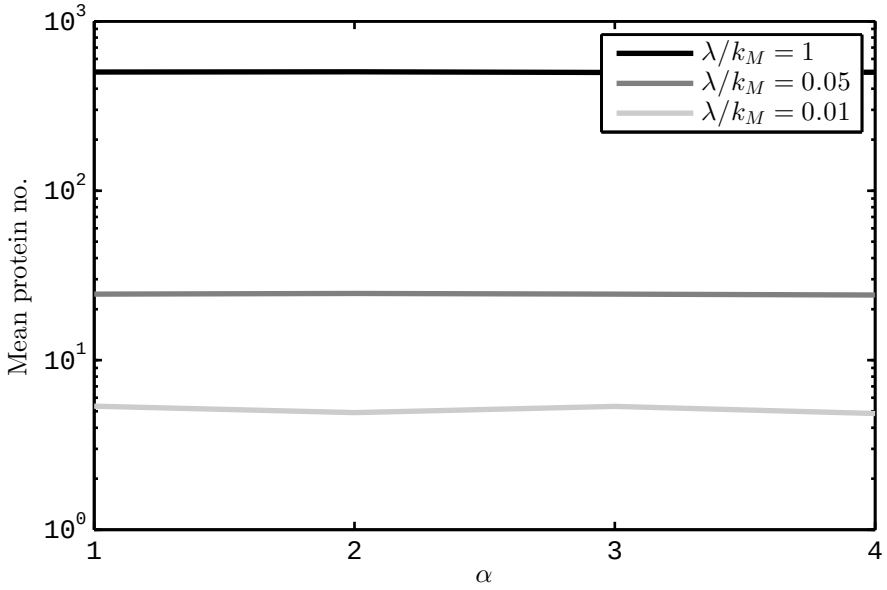

Figure 3.2: **Mean protein level vs rate limiting steps:** the mean protein level is not affected by changing the number of steps but decrease with increasing duration of steps.

number: mRNA Fano factor indicates the dispersion of the distribution, or how spread the probability density of the intervals between mRNA productions and degradations is. For protein noise, we are interested in the variance over mean square (CVS) value, which shows how spread the probability density of the protein number is. CVS is also a good indication of how noisy the regulation of one gene to

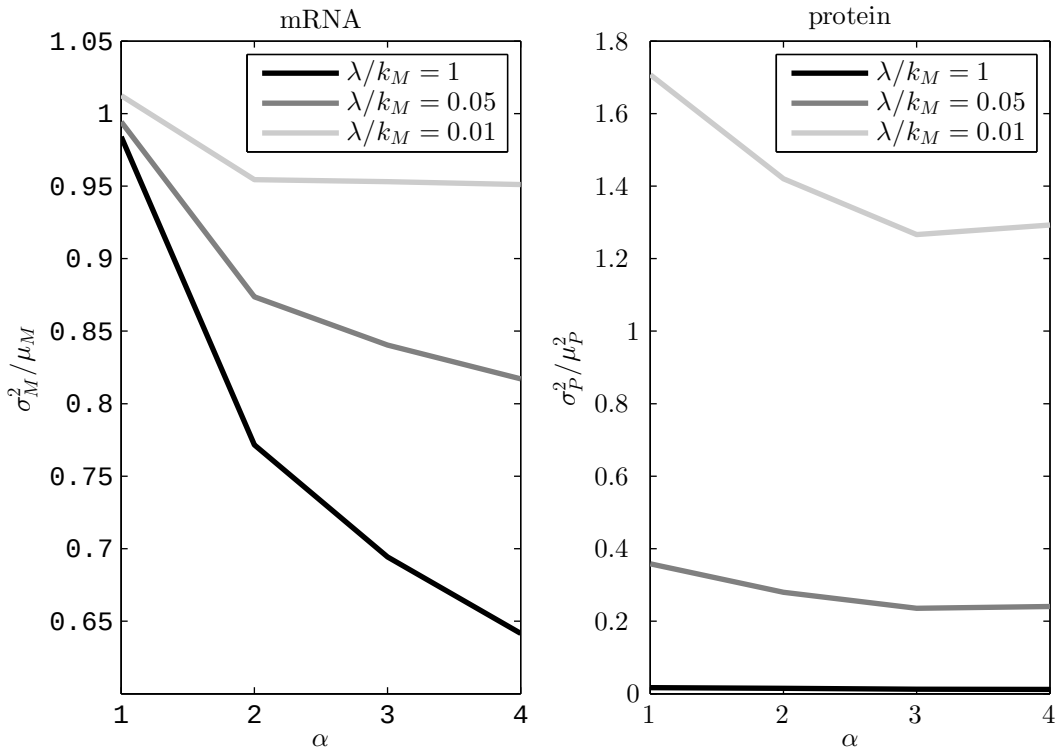

Figure 3.3: **Noise level in mRNA and protein number:** The mRNA noise is indicated by the Fano factor σ^2/μ. The protein noise is indicated by variance over mean square (CVS) value σ^2/μ^2.

another is.

If there is only one rate limiting step in transcription initiation ($\alpha = 1$), the mRNA number follows a Poission distribution with the Fano factor of approximately 1. As we increase the number of steps, there is less noise in transcription initiation, the mRNA level becomes more sub-Poisson resulting in less variance in protein number. Because of the noise in mRNA degradation, even in the limit $\alpha \to \infty$ (the promoter delay reaches constant), there is always noise in mRNA number with the Fano factor's lower bound of 0.5[30]. Due to the first-order translation, the protein number is always super-Poissonian.

The noise level is also dependent of the level of expression: with smaller values of λ/k_M corresponding to lower mRNA and protein levels, the low copy number effect becomes more critical leading to an increase in the Fano factor of mRNA (left figure), the increase in the variance over mean square is even more drastic (right figure).

We should note that despite the noise reduction in mRNA number by increasing the number of steps, the reduction in protein noise only become observable at low expression level, when the mRNA and protein numbers are correlated. Increasing the transcription rate, which makes the system more deterministic, results in a much

more stable expression level. With $\alpha = 4$, $\lambda/k_M = 0.01$, the CVS in protein number is 1.3, greater than that (CVS=0.35) of $\alpha = 1$, $\lambda/k_M = 0.05$.

3.2 Toggle switch

We start by investigating how the behavior of the toggle switch is affected by the kinetics of transcription initiation. The system comprises two genes A and B mutually repressing each other, as shown in figure 2.5. When one gene expresses, its proteins keep the expression of the other gene low and restricts the repression by this gene on itself. Such positive feedback of gene expression leads to the system's multistability[38], which forces gene expressions to stay at the current state, or noisy attractor[32] rather switching to another one.

The 2-gene circuit's stability is classified into two classes: bistability and unistability. The system is said to be bistable if the mean expression levels of the two genes at stationary phase (when the system does not change the state) are different. If the two genes' expressions are equal at stationary phase, the system is unistable. The circuit then loses the capacity to make decisions. For deterministic model, one usually use the number of stable fixed points with zero yielding rates (equation 2.12) to classify the system's stability[10, 23]. However, for stochastic models with intrinsic noise, the deterministic conclusion on the circuit stability does not always apply. Different transcription kinetics can yield different protein dynamics with the same mean molecule number but different in the noise level. Most importantly, the switching events can occur in the stochastic model.

In this section, we quantify the switch stability by calculating the mean time τ_{TS} for the system to switch from one state to another. As the gene product numbers are more stable, the system are expected to flip less frequently, toward not flipping at all: $\tau_{TS} \to \infty$. For unistable system, we easily see that $\tau_{TS} \to 0$.

The regulation of one gene to another is characterized by the disassociation constant $K = 0.05\mu'$. The small value of K ensures a strong level of repression, making bistability easier to achieve[23]. The transcription initiation of two genes are modeled as follow:

$$Pr_A \xrightarrow{\infty} Pr_A(\tau(P_B)) + M_A(\tau(P_B)) \tag{3.1}$$

$$Pr_B \xrightarrow{\infty} Pr_B(\tau(P_A)) + M_B(\tau(P_A)) \tag{3.2}$$

in which $\tau(P) \sim \Gamma(\alpha, \alpha\lambda f(P))$. $f(P)$ is the regulation function in equation 2.42.

In figure 3.4, the switching behavior of the toggle switch is shown. The time series of protein number of the two genes (left subfigure) is filled with separate black and gray sections of different widths, indicating bistability and random switching: when one gene expresses, the other gene is turned off. The protein number distribution

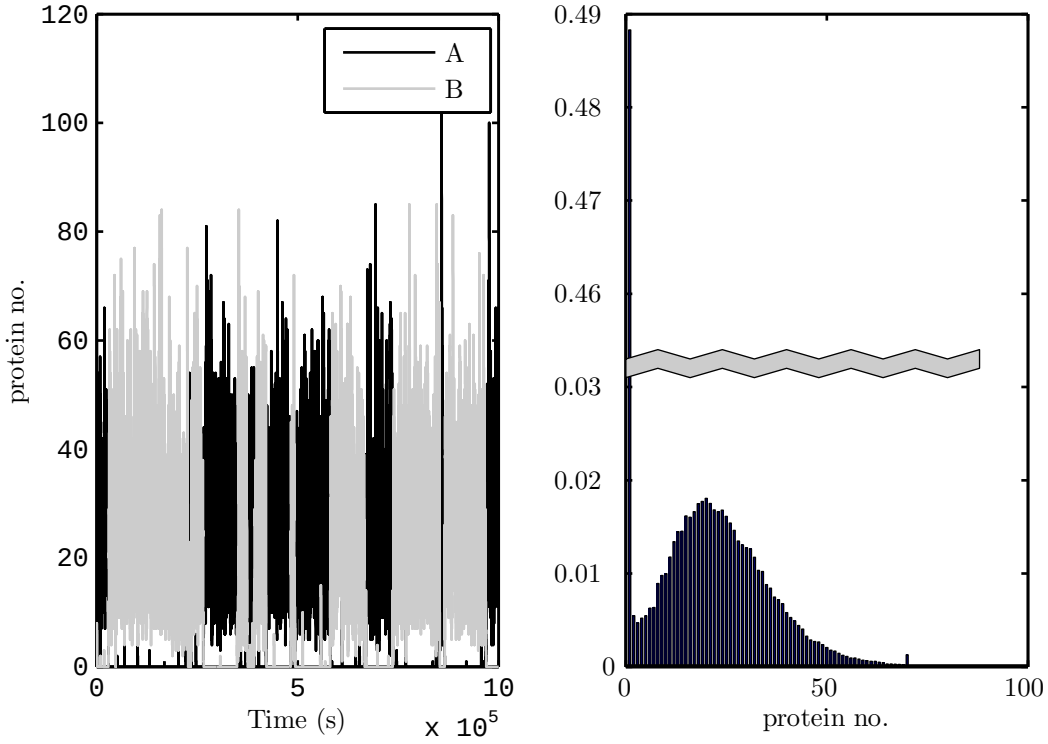

Figure 3.4: **One realization of the switch (left), and probability density of protein numbers (right) with** $\alpha = 4$, $\lambda/k_M = 0.05$

(right subfigure) of each gene is bimodal, with two modes at 0 and 25. The small probability of the protein number between the two modes imply the fast transition from one state to another, once the flipping decision is made.

We define the system's state by comparing the protein levels between the two genes. The system is in state A if the protein number of gene A is greater than that of gene B and vice versa. If the protein numbers of the two genes are equal, the system state is set to its most recent state. From the time series of the system's state, we calculate the distribution of switching intervals (figure 3.5).

We find that a large proportion of intervals have very small values. This is due to the intrinsic noise in gene expression, causing short intervals in the transition phase when the protein numbers of two genes are approximately equal. We filter the intervals for larger values and find that their distribution curve resemble that of exponential distribution. The switching events are therefore random and considered single-step. We estimate the expected value of switching intervals τ_{TS} by fitting its probability density with an exponential curve.

Figure 3.6 shows the stability of the switch with different transcription initiation kinetics. When $\lambda/k_M = 0.1$, the high expression rate of the dominating gene results in a low noise level in its protein number and accordingly the repression on the

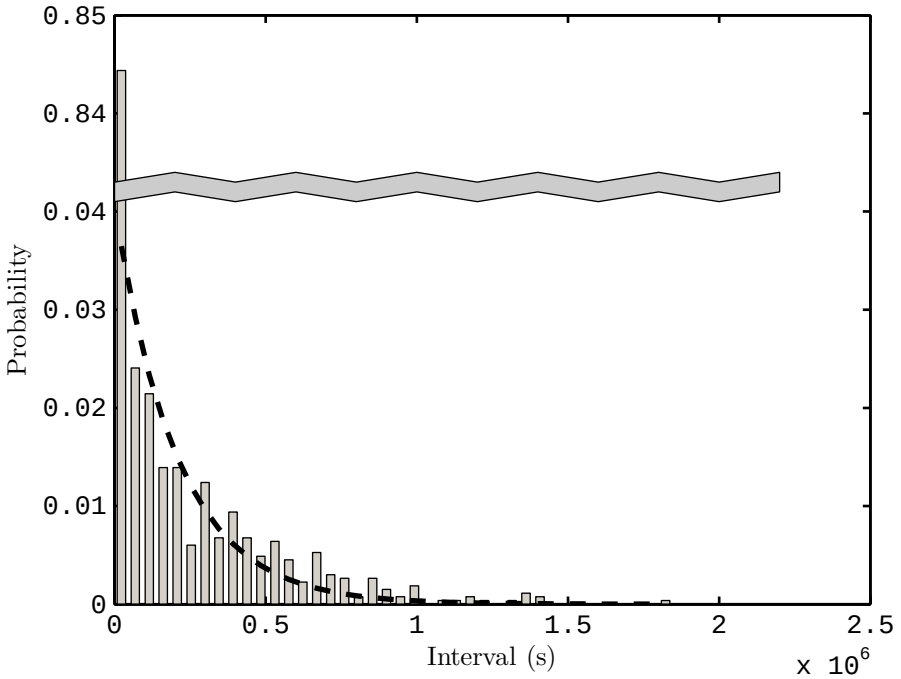

Figure 3.5: **Probability density of the system's switching intervals (τ_{TS}):** the measured density (bar) and the estimation (dash) by fitting the data with exponential distribution

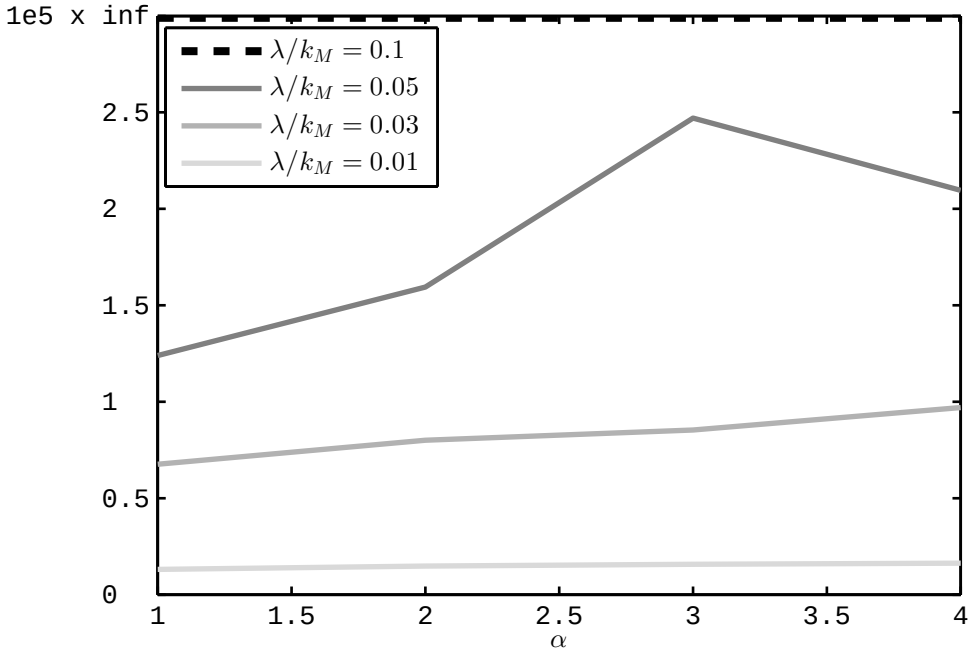

Figure 3.6: **Expected switching interval (τ_{TS}) with different transcription initiation kinetics:** For $\lambda/k_M = 0.1$, $\tau_{TS} \to \infty$, the system never flips.

other gene (from figure 3.3). The system does not change its state during the whole simulation (in 1e7 seconds), thus, the expected switching interval goes to infinity. As

we increase the duration of the steps (by decreasing λ/k_M), τ_{TS} decreases drastically. For λ/k_M as small as 0.01, the low copy number effect makes the repression of the dominating gene on the other one extremely noisy. The system is no longer bistable. The recorded switching events are only due to local noise in the protein number; their intervals are therefore small. Note that, with K proportional to μ', the circuit is still bistable in the deterministic model.

The system loss in stability can be compensated partially by increasing the number of steps in transcription initiation: τ_{TS} increases slightly with increasing values of α. Nevertheless, with α beyond 3, the effect is negligible.

The toggle switch is an autonomous system: the switching behavior is attained with only two genes, without external signals. The motif is found to control many processes critical to cells such as the lysogenic-lytic path decision of cells infected with bacterio-phage lambda[37]. The decision system comprises two genes CI and Cro repressing each other via two promoters λ_{PR} and λ_{PRM}.When The CI gene is activated, the infected cell is in the lysogenic path, still capable of reproduction (along with the viral genetic material) without apparent defections. In normal conditions, the system is known to stay lysogenic as long as 10^8 cell cycles. Under certain changes in the environment, e.g the rise in temperature, the system becomes unstable. When a flipping event occurs, Cro gene expresses, allowing the transcription of lysis genes that destroy the host's membrane and resolve it into viral replications. The motif also plays an important role in establishing cell types, enhancing cell to cell diversity. To determine the duration of the lysogenic path, or the stability of the types, it is necessary to identify the dynamics of the switches controlling the processes. The multiple rate limiting steps characterizing the constituent promoters should be taken into account.

3.3 Amplitude filter

We study the behavior of a filter system, that responses only to input signals within a certain range of amplitude. A genetic motif capable of behaving as a biphasic amplitude filter should trigger the expression of the output gene only for a narrow range of input gene product level, namely its molecular concentration. The region of inputs with active output is referred as the passband, the non-active regions are called stopbands.

The genetic amplitude filter consists of an input gene A with controllable expression, one output gene D and two intermediary genes B and C. Gene A activates the expression of gene B and C; gene B activates the expression of gene D while gene C represses it (figure 3.7). The motifs is used to explain the narrow range of induction triggering the expression of Xbra in Xenopus laevis[7]. With the relative expression level of gene A acting as an input parameter, we investigate the expression dynamics

of gene D with different transcription initiation kinetics of intermediary genes.

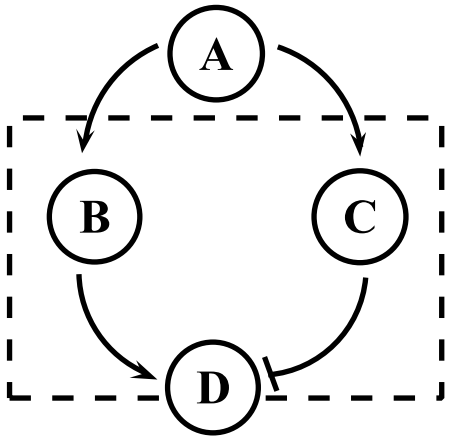

Figure 3.7: **Amplitude filter:** Gene A acts as the input to the filter, gene C and D along with the regulatory connections between genes compose the filter, represented by the dashed box. The protein level of gene D acts as the output.

The transcription dynamics of the gene D is characterized by: $\alpha_D = 2$, $\lambda_D = k_M$. The parameters of the output gene are kept constant to avoid direct effects of rate limiting steps on the protein number at single gene level. We simulate the model with varying shape α' and rate λ' of genes B and C. We set $K_B D = 0.25\mu'$, $K_C D = 0.1\mu'$, in which μ' is the protein level of gene B and C in case of full induction. The difference in the dissociation constants results in a biphasic response. To vary the mean input level, we define the relative expression level of gene A as follow: $p = K_{AB}^{-1} P_A = 0.1 K_{AC}^{-1} P_A \propto P_A$.

For the deterministic model, there is no fluctuation in molecules' numbers. The protein numbers of constituent genes at steady state are:

$$P_B = \mu'(1 + (K_{AB}/P_A)^d)^{-1} = \mu'(1 + p^{-d})^{-1} \tag{3.3}$$

$$P_C = \mu'(1 + (K_{AC}/P_A)^d)^{-1} = \mu'(1 + (10p)^{-d})^{-1} \tag{3.4}$$

$$P_D = \mu(1 + (K_{BD}/P_B)^d)^{-1}(1 + (P_C/K_{CD})^d)^{-1} \tag{3.5}$$

From the above equations, we see that P_D is a direct function of p, independent of μ'. The deterministic results is therefore unaffected by the number of steps α', or their durations $(\alpha\lambda)^{-1}$; it is a good approximation for high expression levels $(\mu' \to \infty)$.

Figure 3.8 shows the regulation functions of gene A directly on gene B and C, and indirectly on D. The functions, taking the relative expression level p of the input gene as an argument, indicate the possibility of the genes in the active state. We can calculate the expected protein level of constituent genes proportional to their activation probability.

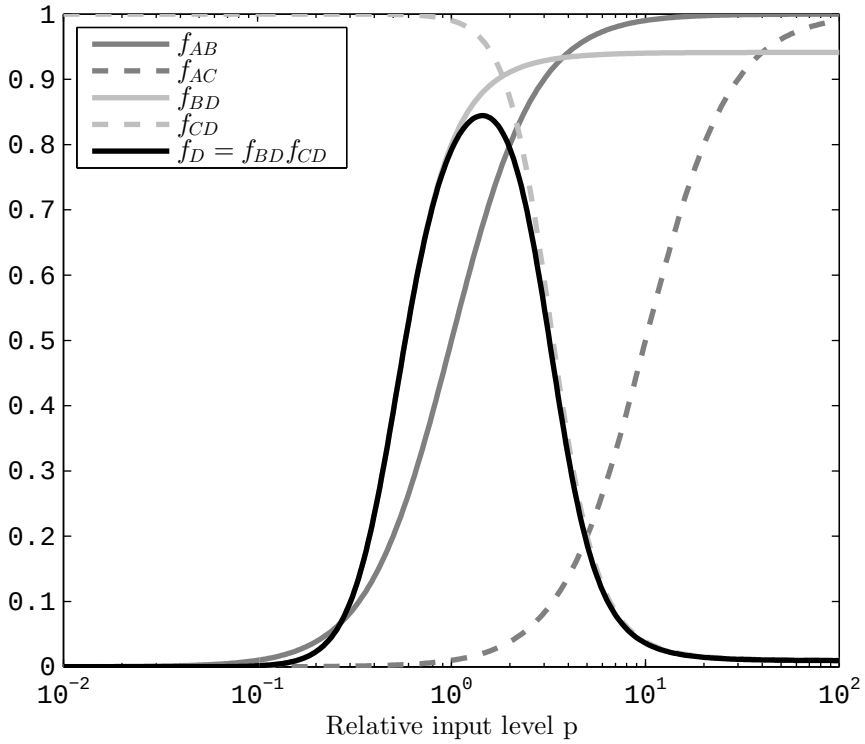

Figure 3.8: **Deterministic responses of the amplitude filter:** f_{AB}, f_{AC} are the probabilities that gene B and gene C are ON. f_{BD}, f_{CD} are the independent regulation functions of gene B and C over gene D. f_D is the probability that gene D is ON.

With very small amount of input proteins ($p \ll 1$), gene B is not activated, and does not produce any proteins. Consequently, there is no output signal as the output gene is not induced. With unrestrained amount of input proteins ($p \gg 1$), the output gene, though induced by gene B's proteins, is repressed completely by the abundance of proteins of gene C. The input regions with very large and small values of p are thus the system stopbands. Only when p is of the same order as 1, the expression of gene D is allowed. The output level is maximized at $p = 1$.

We simulate the stochastic model with the relative input level p ranging from 0.01 to 100, for 10^7 s with the sampling interval of 10 s. We study how much it deviates from the deterministic response with varying values of the shape and rate of transcription. The mean output level of the output gene D is presented in figure 3.9. As expected, the curve from stochastic simulations resembles the curve in figure 3.8, with the output maximized with a specific value of input. Lowering the value of α', which causes more noise in the expression of intermediary genes, results in slight degradation in performance in terms of the response of the filter. That is, the maximum output protein level is lowered, and the transitions between the basspand and the stopbands become less steep. In addition, the increased noise makes the passband to shift toward a higher input level, since the distributions resulting from

the model tend to have right skew.

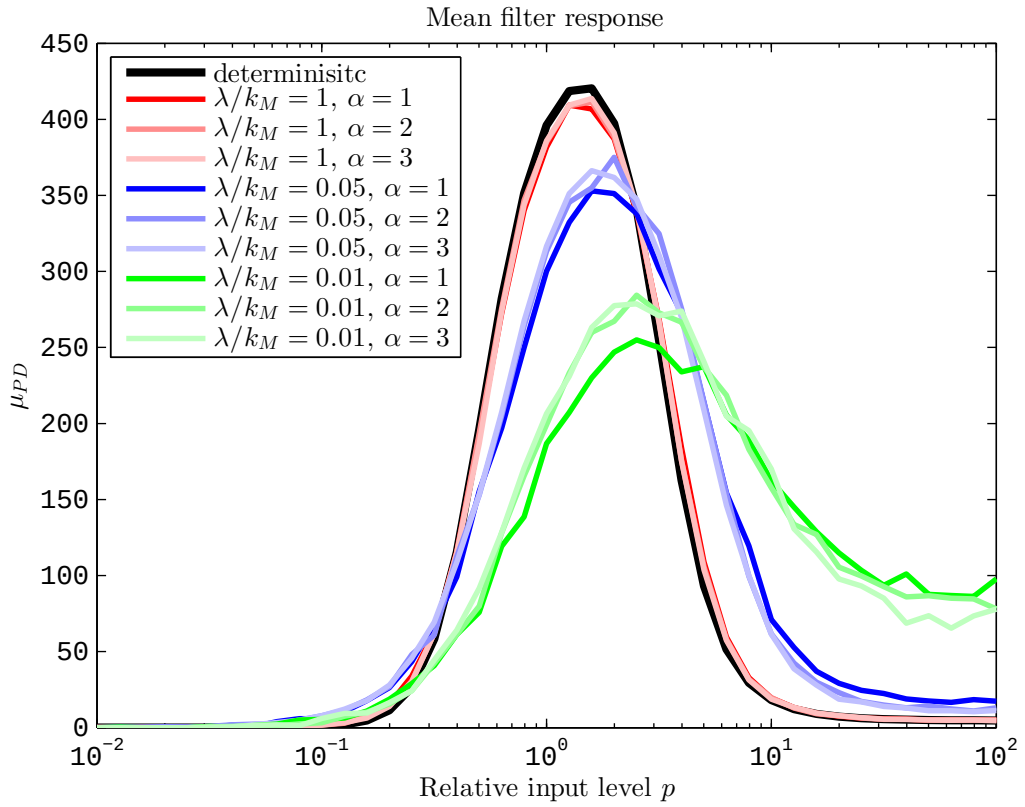

Figure 3.9: **Amplitude filter mean response as a function of relative input level p, in comparison with the deterministic response (black):** The stochastic model is simulated with varying numbers of steps (indicated by the light intensity), and the duration of each step (indicated by the hue) in transcription initiation.

As we vary the mean duration of steps in transcription initiation λ', the mean expression levels μ' of the constituent genes decrease. Noise derived from the low copy number effect propagates to the output gene's transcription initiation and lowers the output product level when the input is in the passband even further, qualitatively similar to the case of lower values of α'. With $\lambda'/k_M = 0.01$, changing α' from 2 to 1 reduces the output peak value by 10%; changing α' from 3 to 1, the reduction is 12%. Figure 3.9 also shows that the change in the filter's performance resulted from adjusting rate-limiting steps' durations is more significant than adjusting the number of steps. With decreasing values of λ'/k_M from 1 to 0.01, $\alpha' = 3$, the maximized protein number of gene D is lowered by 40%. However, on the right stopband, the leakiness in the repression of gene C over gene D occurs as the noise in the repressor number increases. The output gene expression is therefore still allowed at low level even in this stopband.

Note that not only the quality of the filter but also its passband location are affected by the transcription dynamics of gene B and gene C: With longer dura-

tions of rate-limiting steps, the position of the passband is slightly shifted toward higher relative input level. The shift resulted from increasing the number of steps is negligible.

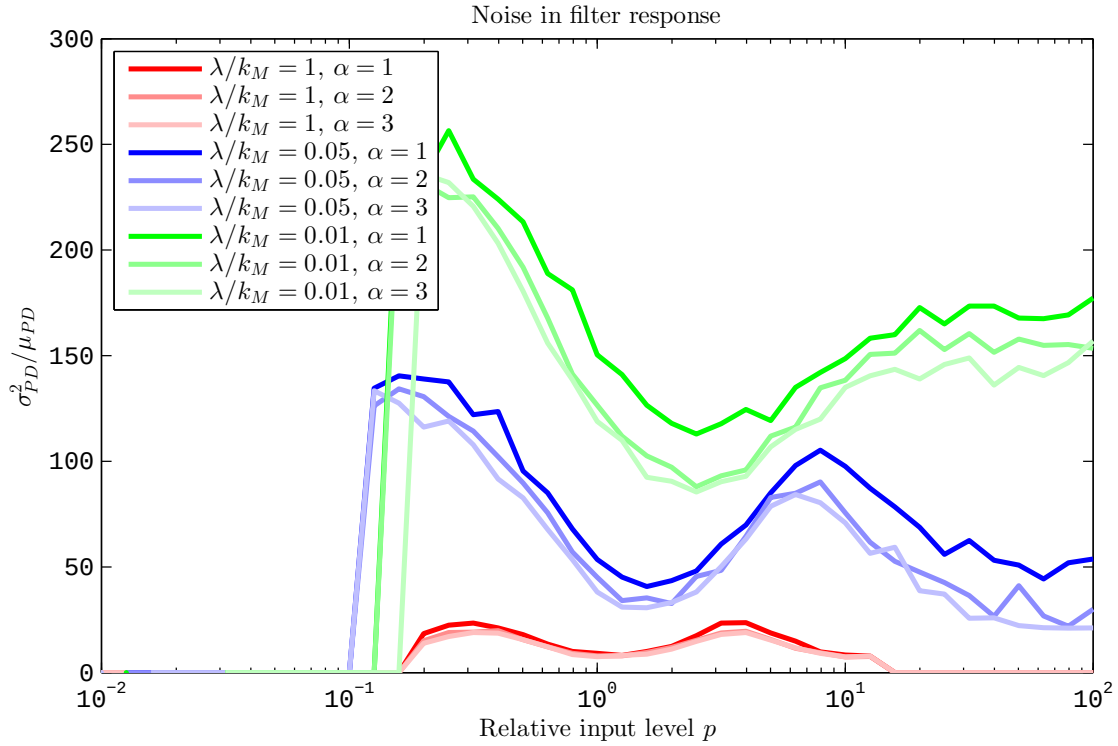

Figure 3.10: **Noise in the response σ^2_P/μ_P of the output gene D as a function of the relative input level p, with different kinetics in transcription initiation:** different levels of brightness denote different shape parameter α'. The changes in the response noise are nonlinear and more drastic with low mean levels.

Due to the stochasticity of our model, we study the fluctuation in the output production level, as a function of the input gene level p. The noise in the output protein is not, as expected from the results on the single gene expression, monotonically dependent on the output level. From figure 3.10, we also observe the noise amplification in the transitions between the passband and stopbands. As λ/k_M decreases, the increasing noise in the number of regulators (P_B and P_C) makes the output protein number unpredictable when p's value is close to $10^{-0.5}$ and $10^{0.5}$. For small values of p, gene D does not express, corresponding to the output protein Fano factor of 1. For large values of p, due to the leakiness in the output gene expression, the level of variability in the second stop band is raised with decreasing λ/k_M.

Figure 3.10 also shows that the increase in number of steps in transcription initiation can suppress the noise in the output gene expression level. For $p = 1$ and $\lambda/k_M = 0.01$, the ouput protein's variance over mean ratio is reduced by 16% when α increases from 1 to 2, and 21% when α increases from 1 to 3. It is notable that the changes from altering α is limited, compared to adjusting the value of λ.

3.4 Frequency filter

In this section, we study another motif that performs filtering in the frequency domain. It is previously known that changes in the transcriptional kinetics can affect the period's duration and robustness of genetic oscillators[22], so we expect those changes affect the performance of filters derived from circuits as well.

3.4.1 3 gene-Repressilator

We first investigate the behavior of a genetic clock comprising 3 genes that represses one another forming a closed loop. The circuit is therefore called repressilator, in which gene A represses gene B, gene B represses gene C and gene C represses gene A (see figure 2.6).

When gene A is on, its products inhibit the expression of gene B, which in turn allows the expression of gene C. As gene C's product reaches a certain level, gene A is repressed, with decreasing protein number. The products of the constituent genes are thus with negative feedback. Also, because of the slow protein dynamics, characterized by the protein degradation rates, the feedback is delayed by a considerable amount of time, leading to oscillation in the protein number of those genes.

In our stochastic model, we have 3 identical genes with the same expression kinetics as described in section 3.1. The disassociation constants are set to $K_{AB} = K_{BC} = K_{CA} = 0.05\mu'$. The three genes repress one another using the same regulation scheme, as in section 3.2. Figure 3.11 shows a time series of the protein numbers P_A, P_B and P_C, which exhibit oscillation, for $\alpha = 4$, $\lambda/k_M = 1$.

For each pair of values (α, λ), we simulate one cell with the simulation time of 10^8 s, and the sampling interval of 50 s. We calculate the autocorrelation of the protein numbers of each constituent gene over time, as a function of lag time. Should the protein numbers oscillate, their autocorrelation function oscillates with the same frequency and consequently, its spectral power is concentrated to the harmonics of $1/T_{ABC}$. By analyzing the oscillation of this function, we can conclude on the oscillation period T_{ABC} in mean and CV^2 of the constituent genes' protein numbers over time. The results are shown in figure 3.12 and 3.13.

Figure 3.13 shows a decay in the circuit's oscillation robustness as the duration and noise of transcription initiation increases. For $\lambda/k_M = 1$, the CV^2 of the recorded periods is ~ 0.05. For $\lambda/k_M = 1$, the CV^2 of the periods is ~ 0.7. Regarding the mean period of the repressilator, we observe that the oscillation frequency is minimized at $\lambda/k_M \sim 10^{-1.4}$. With lower values of λ, the intrinsic fluctuations in protein numbers become the determinant of the circuit's dynamics, resulting to lower values of measured periods. Note that, in this case, proteins of genes are produced slowly in bursts and decay before another burst occurs, which leads to a seemingly

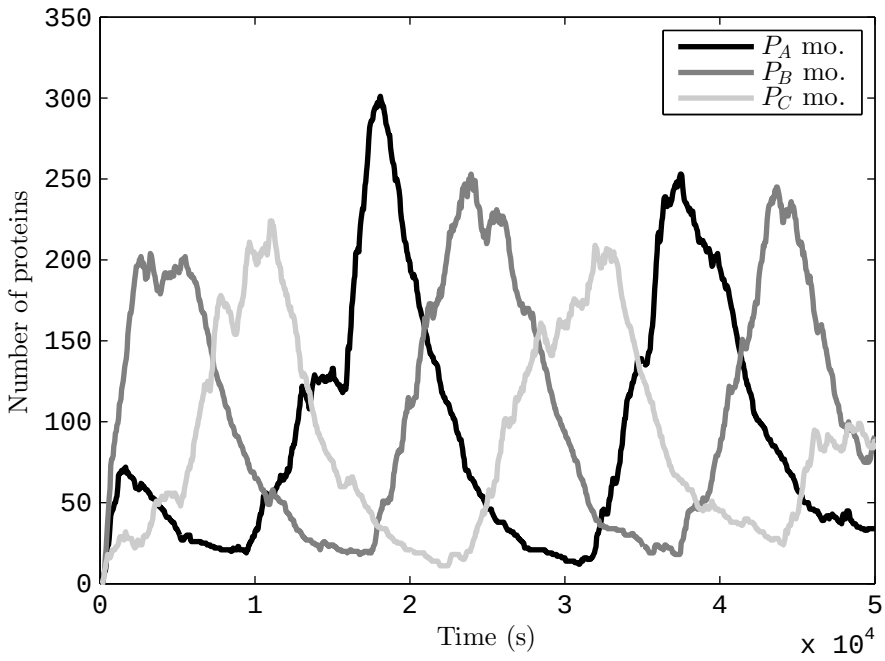

Figure 3.11: **The protein numbers of gene A, B and C in a single cell over time, for $\alpha = 4$, $\lambda/k_M = 1$**

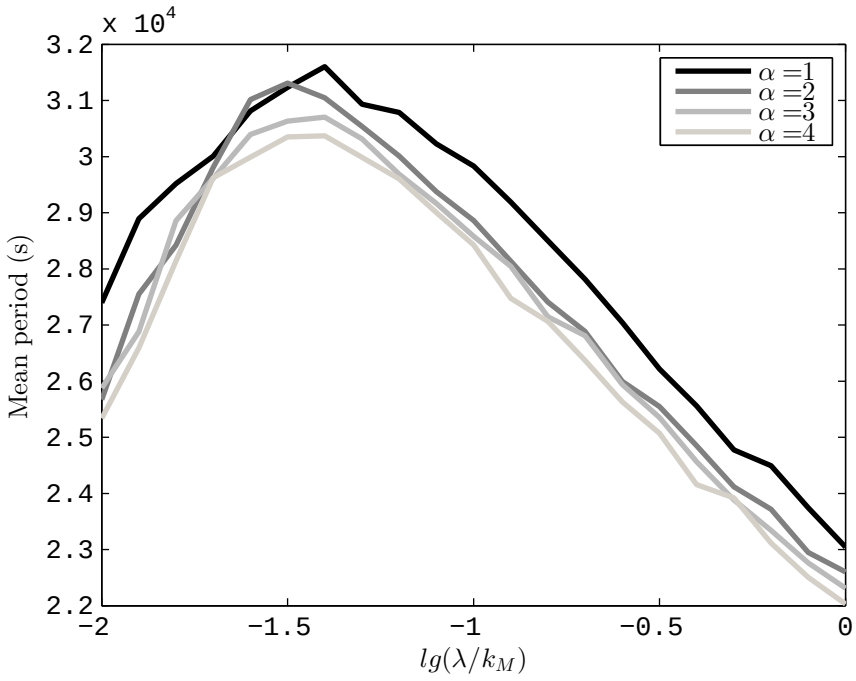

Figure 3.12: **Mean period of the repressilator with different kinetics in transcription initiation:** the intensities of the lines correspond with values of shape parameters α

oscillating behavior. We can consider the range of λ marking the maximum mean period value toward infinity is the functional band of the repressilator.

We also observe that by increasing the number of steps, we can, in a limited

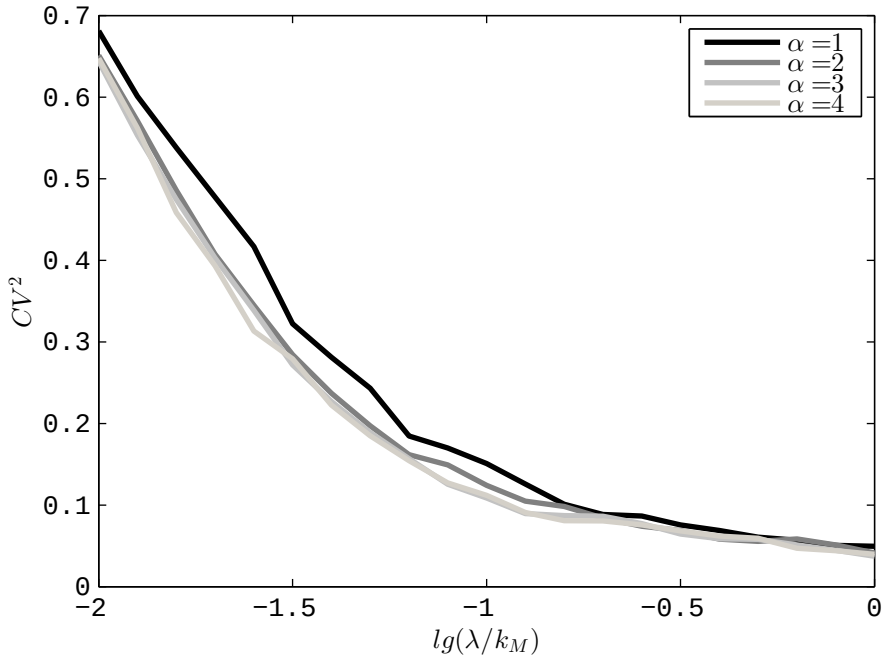

Figure 3.13: CV^2 **of the period**

degree, compensate the loss of robustness in oscillation when lowering the value of λ. From 3.12, we observe a change in the mean period of the circuits as α increases from 1 to 4. Additionally, the functional band of the circuit is widened. For $\alpha = 1$, the circuit only exhibits oscillation with λ/k_M greater than $\sim 10^{-1.3}$; while for $\alpha = 4$, oscillation is established with λ/k_M greater than $\sim 10^{-1.5}$.

3.4.2 Frequency filter with different transcription initiation kinetics

We constructed a motif than can perform a low-pass frequency filtering composed of four genes (A through D). This filter suppresses highly transient signals while letting slowly varying signals to pass though. Such a motif would filter out fast fluctuations in the number of regulatory molecules acting as the filter's input. The structure of the filter is shown in figure 3.14. Gene D is the input gene, required to enable the expression of gene A. Three genes A, B and C repress one another creating the repressilator with the clock behavior as in section 3.4.2.

When a periodic signal P_D is applied, the behavior of this circuit should vary, depending on the frequency of the signal. When the signal is of high frequency, the fluctuation of the input signal does not last long to enough to have significant impact on the expression gene A before the next fluctuation occurs. The feedback loop should be the main responsible for the frequency content of the output. For low frequencies, the input signal will disconnect the feedback loop periodically, and lower

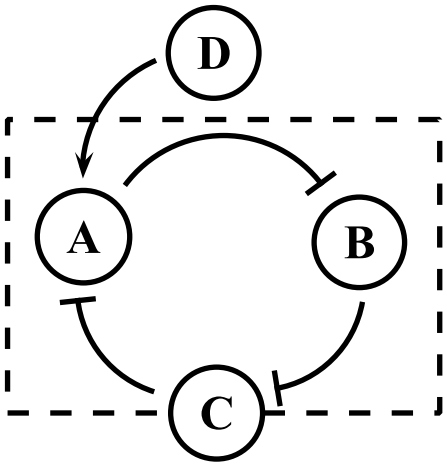

Figure 3.14: **The frequency filter circuit**

frequencies, including that of P_D , are introduced in the output. It is expected that the modulated circuit would have a synchronization point where the input frequency and that of the repressilator are equal. Additionally, the phase in oscillation of the output signal (P_C) is shifted to match the input signal's phase.

To simplify the model, we consider the regulation of the gene D to gene A following a Hill function, with the coefficient $d' \to \infty$. That is, the regulation becomes boolean, with a threshold of K_{DA}. We omit the expression model of gene A, and associate the frequency of the input signal to the frequency of this regulation, denoted by $X \in \mathbb{B}$ switching between 0 and 1. It is notable when X is constant and equal to 1, the output signal P_C oscillates with the repressilator's autonomous period T_{ABC} as described in section 3.4.1.

Next, we apply an unbiased Boolean square wave to X, that is X's value flip between 0 and 1 every fixed interval T/2. We set $X(t) = 0$ for the time t in $[kT, (k + 1/2)T_X)$ with any integer k, and $X(t) = 1$ otherwise. The input signal's period and frequency is therefore T_X and $f_X = 1/T_X$. The autocorrelation function of X is a triangular wave of the same frequency. By using Fourier transform, we find the spectral power density (PSD) of the autocorrelation function, which indicates how much of the signal power per unit frequency is concentrated around certain frequency. Specifically, the PSD of X at frequency f_X is $4\pi^{-2}$.

We measure the power spectral density the output P_C at the frequency f_X, in comparison with that of the input signal X. The spectral power value indicates how effective the input's frequency f_X propagates to the output gene's expression. An example is shown in figure 3.15 , with the input PSD plotted for reference. The circuit's response, corresponding with $\lambda/k_M = 0.05$ indicates a low-pass behavior in the frequency domain. Frequencies lower than those corresponding to the mean period of the three-gene sub-motif when functioning independently (T_{ABC}, see fig-

ure 3.12) are slightly attenuated. Frequencies higher than $1/T_{ABC}$ are filtered out from the PSD of the output autocorrelation function. The transition band is located at frequency $1/T_{ABC}$. As we increase the number of steps in transcription initiation without affecting the total duration of steps, we observe only slight variations in the performance of the filter. The attenuation of the frequency in the passband is reduced, while the positions of the bands remain unaffected.

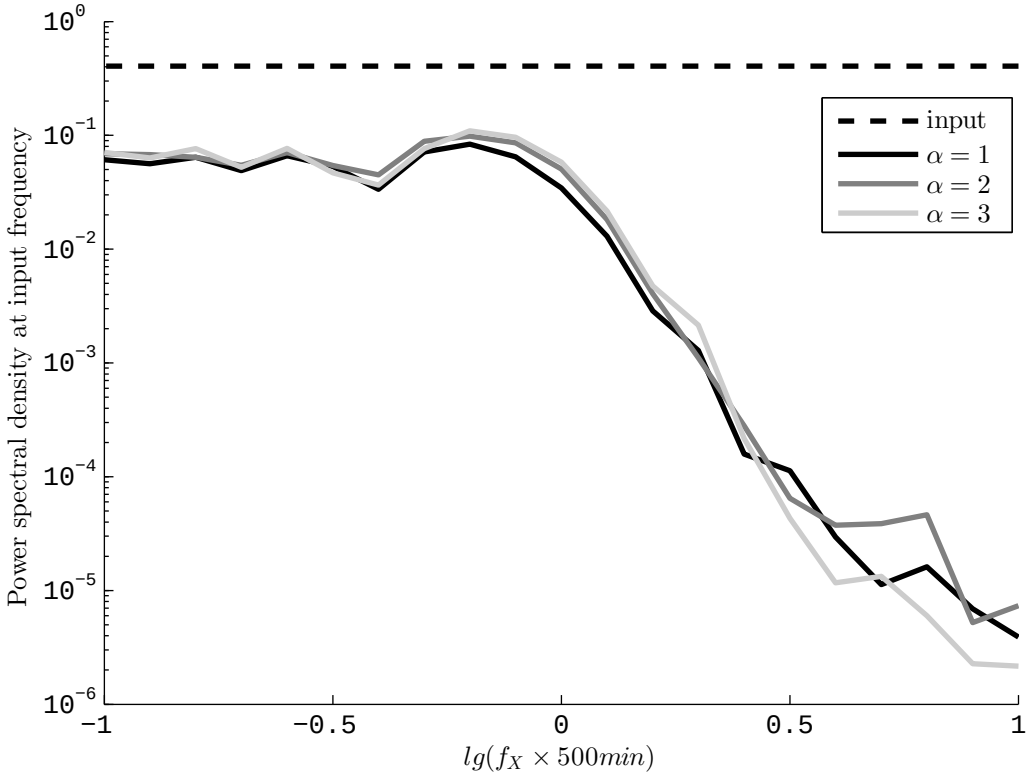

Figure 3.15: **Power spectral density of the frequency filtering motif:** Power spectral density of the frequency filter as a function of the input frequency. Different levels of gray denote different shape parameter α. The simulations were performed with $\lambda/k_M = 0.05$. The dashed black line represents the PSD of the input X at the input frequency, which equal $4\pi^{-2}$.

Next, we vary the duration of steps in transcription initiation with different values of λ. We simulate the model in 10^8 s, with sampling intervals of 60 s, for each set of $< \alpha, \lambda, f_X >$. The result is shown in figure 3.16. With lower transcription rates, the noise in mRNA productions and protein levels are increased, leading to the degradation in the filter performance. The steepness of the transition band is lowered. The output frequency components corresponding to the input frequencies are attenuated drastically even in the pass band. With changing values of λ, the oscillation period of the 3-gene sub-motif when functioning independently varies, resulting in the shift in the position of the filter's cutoff frequency.

Finally, we assessed quantitatively the effects on the output of having different

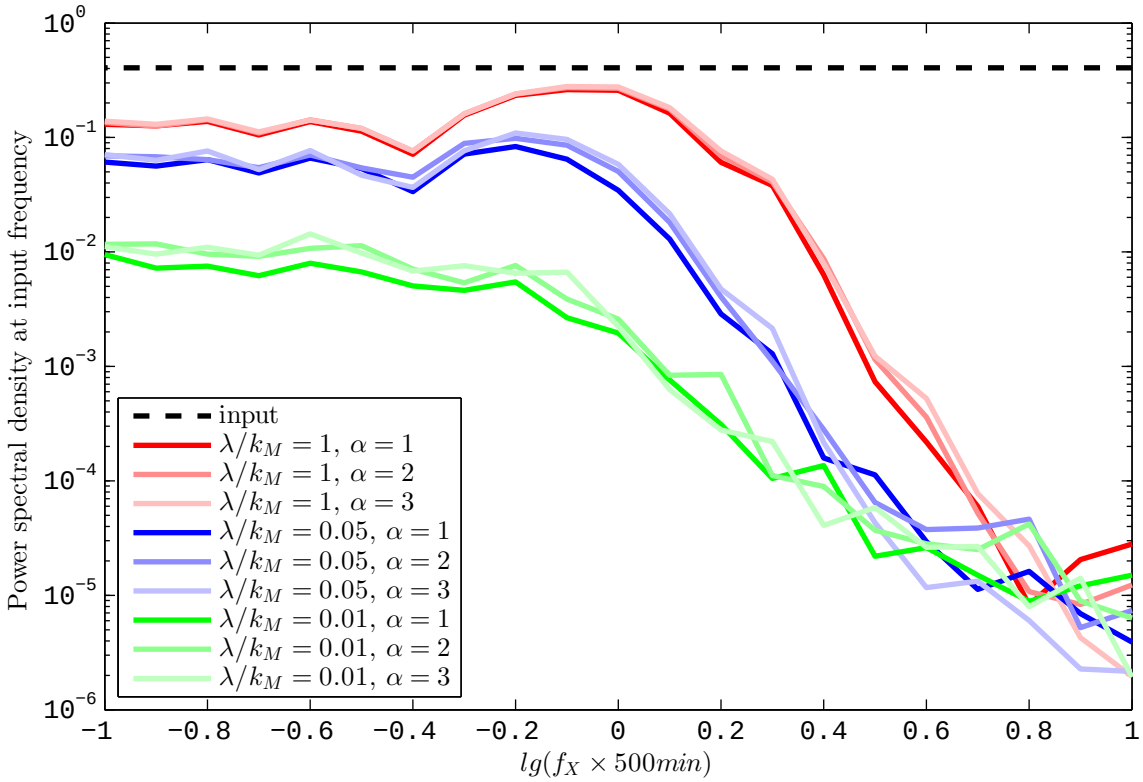

Figure 3.16: **Power spectral density of the frequency filtering motif for various transcription rates:** Power spectral density of the frequency filter as a function of the input frequency, for various shapes α and rates λ of transcription. The changes by varying α are generally nonlinear and more drastic with low mean levels. The dashed black line represents the PSD of the input X at the input frequency, which equal $4\pi^{-2}$.

values of α, for each expression ratio of the input gene. For $\lambda/k_M = 0.01$, increasing α from 1 to 2 causes the magnitude of the PSD in the passband to increase by 236.0%. Increasing α from 1 to 3, increases the PSD by 275.1%. With larger values of λ/k_M, the differences by adjusting α are smaller. In particular, for $\lambda/k_M = 0.05$, the increases are 32.5% and 41.9%. For $\lambda/k_M = 1$, these differences are of the order of 7%.

4. DISCUSSION

In this thesis, we investigated the range of dynamics of a few genetic motifs as a function of the kinetics of transcription initiation of their constituent promoters. We selected three genetic motifs: a toggle switch and two filters in the amplitude and frequency domain; and we studied the behavior of these decision-making and noise-filtering circuits as chemical systems. With the recent evidence on the relevance of the sequence-dependent rate-limiting steps in transcription initiation on RNA numbers, we assessed the performance of these three circuits when making use of different promoter kinetics, which can be characterized by the number and durations of the sequential, rate-limiting steps.

As the dynamics of RNA and protein numbers of real genetic systems are stochastic, and significantly differ from deterministic models, the motifs were expected to function differently from their ideal behavior. By employing the stochastic approach, we built the model of single gene expression, with sequential steps in transcription initiation, and applied it to the circuit models. The sequence-dependence of promoter kinetics is modeled by the time that each sub-process is set to take to be completed. With each set of parameters describing the transcription kinetics of the constituent genes, we simulated the systems using the DSSA and compared the circuits' dynamics with ODE models.

We found that by prolonging the duration of the steps in transcription initiation, the low copy number effect on the expression of constituent genes becomes more drastic, leading to the degradation of the genetic filters' performance. As for the toggle switch, with lower expression rate of the constituent genes, the switching events are more frequent but the discrimination between the system's noisy attractors becomes harder. The multi-step nature of the process, by lowering the noise in mRNA production intervals, can compensate for the low copy noise to some limited extent. We suggest that natural low-expressing genes may employ a multi-step transcription process to achieve higher robustness for their composing motifs' behavior.

We also found that the transcriptional kinetics can affect the characteristics of the filtering circuits. With increasing noise in the expression of the amplitude filter' constituent genes, the location of the filter's pass-bands are shifted toward higher input value. For the frequency filter, the cutoff frequency that separates the pass-band and the stop-band is found to be adjustable, with changing the transcriptional

steps' total duration as "crude tuning". Changing the number of transcriptional steps, though with limited effect in comparison with that on the steps' duration, is non-negligible and can be employed as "fine tuning".

In conclusion, both number and kinetics of steps in transcription initiation are strongly affected by the promoter sequence. Changes in this sequence on the constituent genes of motifs is likely one of the a degrees of freedom of their evolutionary process in natural organisms. This fact also suggests the possibility of construction of synthetic circuits with arbitrarily wired connections.

BIBLIOGRAPHY

[1] A.Arkin, J. Ross, and H. H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.

[2] J. A. Bernstein, A. B. Khodursky, P. Lin, S. Lin-Chao, and S. N. Cohen. Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays. *Proc. Natl. Acad. Sci.*, 99(15):9697–9702, 2002.

[3] D. Bratsun, D. Volfson, L. S. Tsimring, and J. Hasty. Delay-induced stochastic oscillations in gene regulation. *Proc. Natl. Acad. Sci. USA*, 102(41):14593–14598, 2005.

[4] A. Brenicke, A. Marchfelder, and S. Binder. RNA editing. *FEMS Micro. Rev.*, 23:297–316, 1999.

[5] Geoffrey M Cooper. *The Cell: A Molecular Approach.* Sunderland (MA), 2nd edition, 2000. ISBN 0878931066.

[6] P. L. DeHaseth and J. D. Helmann. Open complex formation by *Escherichia coli* RNA polymerase: the mechanism of polymerase-induced strand separation of double helical DNA. *Mol. MicroBiol.*, 16(5):817–824, 1995.

[7] S. Dyson and J. B. Gurdon. The interpretation of position in a morphogen gradient as revealed by occupancy of activin receptors. *Cell*, 93:557–568, 1998.

[8] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(20):335–338, 2000.

[9] B. Albert et. al. *Molecular Biology of the Cell.* 4 edition, 2002. ISBN 0-8153-3218-1.

[10] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli. Nature*, 403(20):339–342, 2000.

[11] Bruck Gibson. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A*, 104:1876–1889, 2000.

[12] D. T. Gillespie. A general method for numerically simulating the stochatic time evolution of coupled chemical reactions. *J. Comp. Phys.*, 22:403–434, 1976.

[13] D. T. Gillespie. Concerning the validity of the stochastic approach to chemical kinetics. *J. Stat.Phys.*, 16(2):311–318, 1977.

[14] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 81:2340–2361, 1977.

[15] I. Golding and E. C. Cox. RNA dynamics in live *Escherichia coli* cells. *Proc. Natl. Acad. Sci. USA*, 101(31):11310–11315, 2004.

[16] S. J. Greive and P. H. Hippel. Thinking quantitatively about transcriptional regulation. 6:221–232, 2005.

[17] P. J. Halling. Do the laws of chemistry apply to living cells? *Trends. Biochem. Sci.*, 14(8):317–318, 1989.

[18] M. Kandhavelu, H. Mannerström, A. Gupta, A. Häkkinen, J. Lloyd-Price, O. Yli-Harja, and A. S. Ribeiro. In vivo kinetics of transcription initiation of the lar promoter in *Escherichia coli*. evidence for a sequential mechanism with two rate-limiting steps. *BMC Sys. Biol.*, 5:149, 2011.

[19] M. Kandhavelu, E. Lihavainen, A. B. Munthukrishnan, O. Yli-Harja, and A. S. Ribeiro. Effects of mg2+ on in vivo transcriptional dynamics of the lar promoter. *Biosys.*, 107:129–134, 2012.

[20] Kroemer Kittel. *Thermal Physics, 2nd edition.* Freeman & Co.: New York, 1980.

[21] T. G. Kurtz. Relationship between stochastic and determinisitc chemical models. 57(7):2976–2978, 1972.

[22] A. Loinger and O. Biham. Stochastic simulations of the repressilator circuit. *Phys. Rev. E*, 76:051917, 2007.

[23] A. Loinger, A. Lipshtat, N. Q. Balaban, and O. Biham. Stochastic simulations of genetic switch systems. *Phys. Rev. E*, 75:021904, 2007.

[24] R. Lutz, T. Lozinski, T. Ellinger, and H. Bujard. Dissecting the functional program of Escherichia coli promoters: the combined mode of action of lac repressor and arac activator. *Nuc. Acid. Res.*, 29(18):3873–3881, 2001.

[25] MATLAB. *version R2012b*. The MathWorks Inc., Natick, Massachusetts, 2012.

[26] W. R. McClure. Rate-limiting steps in RNA chain initiation. *Proc. Natl. Acad. Sci. USA*, 77(10):5634–5638, 1980.

[27] W. R. McClure. Kinetics of open complex formation between *Escherichia coli* RNA polymerase and the *lac-UV5* promoter. evidence for a sequential mechanism involving three steps. *Biochem.*, 24(11):2712–2723, 1985.

[28] Miller. Visualization of bacterial genes in action. *Science*, 169:392–395, 1970.

[29] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298: 824–827, 2002.

[30] J. M. Pedraza and J. Paulsson. Effects of molecular memory and bursting on fluctuation in gene expression. *Science*, 319:339–342, 2008.

[31] I. Potapov, J. Lloyd-Price, O. Yli-Harja, and A. S. Ribeiro. Dynamics of a genetic toggle switch at the nucleotide and codon levels. *Phys. Rev. E.*

[32] A. S. Ribeiro and S. A. Kauffman. Noisy attractors and ergodic sets in models of gene regulatory networks. *J. Theor. Biol.*, 247(4):743–755, 2007.

[33] A. S. Ribeiro and J. Lloyd-Price. SGN Sim, a stochastic genetic networks simulator. *Bioinfo.*, 23(26):777–779, 2007.

[34] A. S. Ribeiro, R. Zhu, and S. A. Kauffman. A general modeling strategy for gene regulatory networks with stochastic dynamics. *J. Comp. Biol.*, 13(9): 1630–1639, 2006.

[35] M. R. Roussel and Rui Zhu. Validation of an algorithm for delay stochastic simulation of transcription and translation in prokaryotic gene expression. *Phys. Biol.*, 3:274–284, 2006.

[36] R. M. Saecker, Jr. M. T. Record, and P. L. Dehaseth. Mechanism of bacterial transcription initiation: Promoter binding, isomerization to initiation-competent open complexes, and initiation of RNA synthesis. *J. Mol. Biol.*, 412 (5):754–771, 2011.

[37] M. A. Shea and G. K. Ackersf. The 0_R control system of bacteriophage lambda: A physical-chemical model for gene regulation. *J. Mol. Biol.*, 181:211–230, 1985.

[38] W. K. Smits, O. P. Kuipers, and J. Veening. Phenotypic variation in bacteria: the role of feedback regulation. *Nat. Rev. Microbiol.*, 4(4):259–271, 2006.

[39] Y. Taniguchi, P. J. Choi, G. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[40] Z. Wang and J. Zhang. Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proc. Nat. Acad. Sci.*, 108(16):E67–E76, 2011.

[41] D. M. Wolf and A. P. Arkin. Motifs, modules and games in bacteria. *Curr. Opin. in Microbio.*, 6(2):125–134, 2003.

[42] K. Xia, M. Manning, H. Hesham, Q. Lin, C. Bystroff, and W. Colon. Identifying the subproteome of kinetically stable proteins via diagonal 2D SDS/PAGE. *Proc. Natl. Acad. Sci. USA*, 104(44):17329–17334, 2007.

[43] R. Zhu, A. S. Ribeiro, D. Salahub, and S. A. Kauffman. Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *J. Theo. Biol.*, 246:725–745, 2007.